

Estimation and Prediction in Transformed Nested Error Regression Models

Shonosuke Sugasawa* and Tatsuya Kubokawa†

University of Tokyo

Abstract

For analyzing positive or bounded data, this paper suggests parametrically transformed nested error regression models (TNERM), which transform the data flexibly to follow the normal linear mixed regression. As useful transformations, we consider the dual power transformation for positive data and the dual power logistic transformation newly proposed for bounded data. We provide a procedure for estimating consistently the parameters of the proposed model and a predictor based on the consistent estimators. Then, in order to calibrate uncertainty of the transformed empirical best linear unbiased predictor, we derive both unconditional and conditional prediction intervals with second-order accuracy based on the parametric bootstrap method. The proposed methods are investigated through simulation and empirical studies.

Key words and phrases: dual power transformation, linear mixed model, nested error regression model, parametric bootstrap, prediction intervals, small area estimation.

1 Introduction

The linear mixed models with both random and fixed effects have been extensively and actively studied in recent years from both theoretical and applied aspects in the literature. As specific normal linear mixed models, the Fay-Herriot model (Fay and Herriot, 1979) and the nested error regression models (Battese, Harter and Fuller, 1988) have been used in small-area estimation since direct estimates like sample means for small areas have unacceptable estimation errors because of small sample sizes in small areas. The model-based shrinkage methods such as the empirical best linear unbiased predictor (EBLUP) are very useful for providing reliable estimates for small-areas with higher precisions by borrowing data in the surrounding areas. Recently, several approaches for small area estimation are proposed and investigated in terms of both parametric and nonparametric aspects. For example, see Hall and Maiti (2006a, b), Chambers et al.(2014), Chaudhuri and Ghosh (2011), Jiang and Nguyen (2012) and Opsomer et al.(2008). Also see Ghosh and Rao (1994), Rao (2003), Datta and Ghosh (2012) and Pfeiffermann (2013) for a good survey on this topic.

*Graduate School of Economics, University of Tokyo, E-Mail: shonosuke622@gmail.com

†Faculty of Economics, University of Tokyo, E-Mail: tatsuya@e.u-tokyo.ac.jp

In the most of literatures related to the Fay-Herriot model or the nested error regression model, it is assumed that the data has real values. However, we are often faced with the data with positive values like the price data and fitting such data to the normal distributions in nested error regression model may be inappropriate in the case that distribution of data is skewed. Moreover, if we analyze bounded data like ratio data as real-valued data, we may cause the problem that prediction intervals of model-based predictors may be out from the range of the bounded data. Thus, this paper is focused on developing a methodology for analyzing positive or bounded data by incorporating a parametric transformation into the conventional nested error regression model.

A standard approach to analyzing positive data is to apply the log-transformation to the data, which was investigated in Slud and Maiti (2006) in the nested error regression model. However, such a specific transformation is not always appropriate, and we want to adjust the transformation flexibly to fit the transformed data to a normal linear mixed regression. A conventional method in this direction is the Box-Cox transformation suggested by Box and Cox (1964), described by

$$h^{\text{BC}}(x, \lambda) = \begin{cases} (x^\lambda - 1)/\lambda, & \lambda \neq 0, \\ \log x, & \lambda = 0. \end{cases}$$

However, it is known that the maximum likelihood estimator of the transformation parameter λ in the Box-Cox transformation is not consistent, so that the resulting EBLUP is not consistent to the best linear unbiased predictor (BLUP). Thus, the inconsistency in the estimation of λ is a crucial problem in the context of small area prediction. As an alternative transformation, in this paper, we use the dual power transformation (DPT) (Yang, 2006) which will be described in Section 2. This is a transformation from positive numbers to real numbers, and it can be expected that the maximum likelihood estimator of λ is consistent. For analyzing bounded data, we propose the dual power logistic transformation (DPLT) which transforms the bounded data to real-valued data. This transformation will be also described in Section 2.

Utilizing these transformations, in this paper, we suggest the parametrically transformed nested error regression model (TNERM) defined as

$$h(y_{ij}, \lambda) = \mathbf{x}'_{ij}\boldsymbol{\beta} + v_i + \varepsilon_{ij}, \quad i = 1, \dots, m, \quad j = 1, \dots, n_i,$$

where v_i 's and ε_{ij} 's are mutually independently normally distributed and $\mathbf{x}'_{ij}\boldsymbol{\beta}$ is a fixed effect. Here, $h(\cdot, \lambda)$ is a general transformation function which is characterized with the transformation parameter λ . The examples of $h(\cdot, \lambda)$ treated in this paper are DPT and DOLT. The transformation parameter λ can be used for adjustment, and the proposed model enables us to flexibly analyze the small-area positive or bounded data. The detailed model description is given in Section 2.

In the conventional nested error regression, we predict the quantity $\xi_i = \bar{\mathbf{x}}'_i\boldsymbol{\beta} + v_i$, where $\bar{\mathbf{x}}_i = \sum_{j=1}^{n_i} \mathbf{x}_{ij}$ by the empirical best linear unbiased predictor $\hat{\xi}_i^{\text{EB}}$. Sugawara and Kubokawa (2014) proposed the parametrically transformed Fay-Herriot model and the gave second-order unbiased estimators of MSE of $\hat{\xi}_i^{\text{EB}}$. However, in the transformed model, the quantity of interest is the inversely transformed function $h^{-1}(\xi_i, \lambda)$ rather than ξ_i . Thus, in this paper, we consider to predict $h^{-1}(\xi_i, \lambda)$ and propose the transformed empirical best linear unbiased estimator (TEBLUP), namely $h^{-1}(\hat{\xi}_i^{\text{EB}}, \hat{\lambda})$. Since this predictor is expected to give reliable predicted

values for small-areas with higher precisions, it is important to assess uncertainty of TEBLUP. In the context of small area estimation, we have two approaches to measuring the uncertainty: One is to provide an estimate of the mean squared error (MSE) of the prediction (see Datta and Lahiri, 2000, Datta et al., 2005 and Hall and Maiti, 2006a,b), and the other is the prediction (confidence) intervals (see Basu, et al., 2003, Chatterjee et al., 2008, Datta et al., 2002, Diao et al., 2014 and Yoshimori and Lahiri, 2014).

The goal of this paper is to construct prediction intervals of $h^{-1}(\xi_i, \lambda)$ based on $h^{-1}(\hat{\xi}_i^{\text{EB}}, \hat{\lambda})$. Since it is harder to derive an analytical prediction interval with suitable accuracy based on the Taylor series expansion, we here provide a prediction interval with second-order accuracy based on the parametric bootstrap along the line given in Chatterjee, Lahiri and Li (2008). We also provide a conditional prediction interval given data in the area of interest, motivated from the results of Booth and Hobert (1998) who discussed a conditional MSE and its estimation in generalized linear mixed models.

The paper is organized as follows: In Section 2, we suggest the parametric transformed nested error regression model with DPT or DPLT. Some consistent estimators of parameters in TNERM are also given. In Section 3, we introduce the transformed empirical best linear unbiased predictor and construct unconditional and conditional prediction intervals with second order accuracy based on the parametric bootstrap method. In Section 4, we conduct simulation studies. In Section 5, we apply the proposed model to two data set, the survey data in Japan and crop areas data given in Battese et al.(1988). The concluding remarks are given in Section 6. All the technical proofs are given in the Appendix.

2 Transformed Nested Error Regression Models

2.1 Settings and transformations

Consider the two-stage cluster sampling, namely, m clusters are randomly selected, and data are randomly selected from each selected cluster. For $i = 1, \dots, m$, a random sample taken from the i -th cluster with size n_i is denoted by y_{i1}, \dots, y_{in_i} . The most useful model for analyzing such data is the nested error regression model (NERM) described by

$$y_{ij} = \mathbf{x}'_{ij}\boldsymbol{\beta} + v_i + \varepsilon_{ij}, \quad i = 1, \dots, m, \quad j = 1, \dots, n_i, \quad (1)$$

where v_i 's and ε_{ij} 's are mutually independently distributed as $v_i \sim \mathcal{N}(0, \sigma_v^2)$ and $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma_e^2)$. Here, a vector \mathbf{x}' denotes the transpose of \mathbf{x} , $\mathcal{N}(\mu, \sigma^2)$ denotes a normal distribution with mean μ and variance σ^2 , \mathbf{x}_{ij} is a p -dimensional known covariate associated with y_{ij} , $\boldsymbol{\beta}$ is a p -dimensional unknown vector of regression coefficients, and σ_v^2 and σ_e^2 are unknown components of variance, called 'between' and 'within' components, respectively.

The model (1) is a linear mixed model which incorporates both fixed and random effects, and it has been used for analyzing unit level data in the framework of small-area estimation. When y_{ij} 's are real-valued data, the model (1) is reasonable. However, it is not necessarily appropriate when values of y_{ij} 's are limited to spaces of positive or bounded numbers. Then, we need to consider a transformation of y_{ij} to fit into NERM. In this paper, we consider two types of transformations for the data limited to \mathbb{R}_+ or $\mathbb{R}_{(0,1)}$, where $\mathbb{R}_+ = \{x \in \mathbb{R}; x > 0\}$ and $\mathbb{R}_{(a,b)} = \{x \in \mathbb{R}; a < x < b\}$ for the real space \mathbb{R} . For \mathbb{R}_+ , namely positive data, we use the

dual power transformation (DPT) suggested by Yang (2006), described as

$$h^{\text{DP}}(x, \lambda) = \begin{cases} (x^\lambda - x^{-\lambda})/2\lambda, & \lambda > 0, \\ \log x, & \lambda = 0. \end{cases} \quad (2)$$

for $x > 0$. It is noted that for $z = h^{\text{DP}}(x, \lambda)$, the inverse transformation is expressed as

$$x = \left(\lambda z + \sqrt{\lambda^2 z^2 + 1} \right)^{1/\lambda}$$

for $\lambda \neq 0$, and $x = e^z$ for $\lambda = 0$. When data are restricted on the space $\{x \in \mathbb{R} | x > a\}$ for $a \in \mathbb{R}$, DPT can be extended to $h^{\text{DP}}(x - a, \lambda)$ for analyzing data on the space. For $\mathbb{R}_{(0,1)}$, we newly suggest the dual power logistic transformation (DPLT) given by

$$h^{\text{DPL}}(x, \lambda) = \begin{cases} \left\{ \left(\frac{x}{1-x} \right)^\lambda - \left(\frac{1-x}{x} \right)^\lambda \right\} / 2\lambda, & \lambda > 0, \\ \log \left(\frac{x}{1-x} \right), & \lambda = 0. \end{cases} \quad (3)$$

for $0 < x < 1$. Using the expression of the inverse transformation of DPT, one gets the inverse transformation of DPLT, given by

$$x = \frac{\left(\lambda z + \sqrt{\lambda^2 z^2 + 1} \right)^{1/\lambda}}{1 + \left(\lambda z + \sqrt{\lambda^2 z^2 + 1} \right)^{1/\lambda}}$$

for $\lambda \neq 0$, and $x = e^z / (1 + e^z)$ for $\lambda = 0$. When data are restricted on the interval (a, b) for fixed values a and b , ($a < b$), DPLT can be extended to $h^{\text{DPL}}((x - a)/(b - a), \lambda)$, since $(x - a)/(b - a)$ lies in $(0, 1)$. Thus, we can analyze data on (a, b) using $h^{\text{DPL}}((x - a)/(b - a), \lambda)$.

Let $h(\cdot, \lambda)$ be $h^{\text{DP}}(x, \lambda)$ or $h^{\text{DPL}}(x, \lambda)$ for given λ . The transformation parameter λ is adjusted so that transformed data $h(y_{ij}, \lambda)$'s can fit into NERM. Thus, we can suggest the parametrically transformed nested error regression model (TNERM)

$$h(y_{ij}, \lambda) = \mathbf{x}'_{ij} \boldsymbol{\beta} + v_i + \varepsilon_{ij}, \quad i = 1, \dots, m, \quad j = 1, \dots, n_i. \quad (4)$$

It may be convenient to write the model (4) in matricial forms. Let $\mathbf{y} = (y_{i1}, \dots, y_{in_i})'$, $\mathbf{X}_i = (\mathbf{x}_{i1}, \dots, \mathbf{x}_{in_i})'$, $\boldsymbol{\epsilon}_i = (\varepsilon_{i1}, \dots, \varepsilon_{in_i})'$ and $\mathbf{j}_{n_i} = (1, \dots, 1) \in \mathbb{R}^{n_i}$. Also, define $\mathbf{h}(\mathbf{y}_i, \lambda)$ by

$$\mathbf{h}(\mathbf{y}_i, \lambda) = (h(y_{i1}, \lambda), \dots, h(y_{in_i}, \lambda))'.$$

Then, the model (4) is expressed as

$$\mathbf{h}(\mathbf{y}_i, \lambda) = \mathbf{X}_i \boldsymbol{\beta} + \mathbf{j}_{n_i} v_i + \boldsymbol{\epsilon}_i, \quad i = 1, \dots, m, \quad (5)$$

and $\mathbf{h}(\mathbf{y}_i, \lambda)$ has an n_i -variate normal distribution $\mathcal{N}_{n_i}(\mathbf{X}_i \boldsymbol{\beta}, \sigma_e^2 \mathbf{V}_i(\rho))$ where $\mathbf{V}_i(\rho) = \mathbf{I}_{n_i} + \rho \mathbf{J}_{n_i}$ for $\rho = \sigma_v^2 / \sigma_e^2$, the $n_i \times n_i$ identity matrix \mathbf{I}_{n_i} and $\mathbf{J}_{n_i} = \mathbf{j}_{n_i} \mathbf{j}'_{n_i}$. It is noted that the covariance of $\mathbf{h}(\mathbf{y}_i, \lambda)$ has the intra-class correlation structure, namely $h(y_{i1}, \lambda), \dots, h(y_{in_i}, \lambda)$ are not mutually

independent when $\rho \neq 0$. Let $N = \sum_{i=1}^m n_i$. All the data \mathbf{y}_i 's are described as the N -dimensional vector $\mathbf{Y} = (\mathbf{y}'_1, \dots, \mathbf{y}'_m)'$. Then the joint density function \mathbf{Y} is expressed as

$$f(\mathbf{Y}) = (2\pi)^{-N/2} \sigma_e^{-m} \prod_{i=1}^m \det(\mathbf{V}_i(\rho)) \prod_{i=1}^m \prod_{j=1}^{n_i} h_x(y_{ij}, \lambda) \\ \times \exp \left\{ -\frac{1}{2} \sigma_e^{-2} \sum_{i=1}^m (\mathbf{h}(\mathbf{y}_i, \lambda) - \mathbf{X}_i \boldsymbol{\beta})' \mathbf{V}_i(\rho)^{-1} (\mathbf{h}(\mathbf{y}_i, \lambda) - \mathbf{X}_i \boldsymbol{\beta}) \right\}, \quad (6)$$

where $\prod_{i=1}^m \prod_{j=1}^{n_i} h_x(y_{ij}, \lambda)$ is the Jacobian of the transformation for $h_x(x, \lambda) = \partial h(x, \lambda) / \partial x$. This expression will be used for estimating the unknown parameters $\boldsymbol{\beta}$, σ_v^2 , σ_e^2 and λ .

2.2 Consistent estimators of the parameters

We here provide consistent estimators of the unknown parameters $\boldsymbol{\beta}$, σ_v^2 , σ_e^2 and λ . To this end, we begin by estimating $\boldsymbol{\beta}$, σ_v^2 and σ_e^2 in the case of known λ . In this case, the conventional estimators given in the literature for NERM (1) can be inherited to the transformed model.

Concerning estimation of $\boldsymbol{\beta}$, the maximum likelihood (ML) or generalized least square (GLS) estimator of $\boldsymbol{\beta}$ for known σ_v^2 , σ_e^2 and λ is

$$\hat{\boldsymbol{\beta}}(\rho, \lambda) = \left(\sum_{i=1}^m \frac{n_i \bar{\mathbf{x}}_i \bar{\mathbf{x}}_i'}{1 + n_i \rho} + \sum_{i=1}^m \sum_{j=1}^{n_i} \mathbf{x}_{ij} \mathbf{x}_{ij}' \right)^{-1} \left(\sum_{i=1}^m \frac{n_i \bar{\mathbf{x}}_i z_i(\lambda)}{1 + n_i \rho} + \sum_{i=1}^m \sum_{j=1}^{n_i} \mathbf{x}_{ij} h(y_{ij}, \lambda) \right), \quad (7)$$

where $\rho = \sigma_v^2 / \sigma_e^2$, $\bar{\mathbf{x}}_i = \sum_{j=1}^{n_i} \mathbf{x}_{ij} / n_i$ is the mean of covariates \mathbf{x}_{ij} 's for the i -th area, and

$$z_i(\lambda) = \frac{1}{n_i} \sum_{j=1}^{n_i} h(y_{ij}, \lambda), \quad i = 1, \dots, m,$$

is the mean of the transformed observations. Since $\hat{\boldsymbol{\beta}}(\rho, \lambda) \sim \mathcal{N}_p(\boldsymbol{\beta}, \{\sum_{j=1}^m (n_j \sigma_v^2 + \sigma_e^2)^{-1} n_j \bar{\mathbf{x}}_j \bar{\mathbf{x}}_j'\}^{-1})$, it is clear that $\hat{\boldsymbol{\beta}}(\rho, \lambda)$ is consistent and $\hat{\boldsymbol{\beta}}(\rho, \lambda) - \boldsymbol{\beta} = \mathbf{O}_p(m^{-1/2})$ under the following assumption:

Assumption 1. *The following are assumed for $\bar{\mathbf{x}}_i$ and n_i :*

- (A.1) $m^{-1} \sum_{i=1}^m \bar{\mathbf{x}}_i \bar{\mathbf{x}}_i'$ converges to a positive definite matrix as $m \rightarrow \infty$.
- (A.2) There exist integers \underline{n} and \bar{n} which are positive and independent of m such that $\underline{n} \leq n_i \leq \bar{n}$ for $i = 1, \dots, m$.

Since σ_v^2 and σ_e^2 are unknown, we estimate them and substitute their estimators into $\hat{\boldsymbol{\beta}}(\rho, \lambda)$. In NERM (1) with known λ , for estimation of σ_v^2 and σ_e^2 , the Prasad-Rao estimator, the maximum likelihood (ML) and the restricted maximum likelihood (REML) estimators have been used in the literature, and it would be plausible that those estimators can be used still in TNERM (4) by replacing λ with an estimator. We here clarify conditions that estimators of σ_v^2 and σ_e^2 should satisfy in order to derive prediction intervals given in this paper. For notational convenience, $\mathbf{O}_p(a_n)$ means that every component in $\mathbf{O}_p(a_n)$ is of order $\mathbf{O}_p(a_n)$, and the notation $\mathbf{O}(a_n)$ is defined similarly.

Assumption 2. Let $\hat{\boldsymbol{\sigma}}^2(\lambda) = (\hat{\sigma}_e^2(\lambda), \hat{\sigma}_v^2(\lambda))'$ be an estimator of $\boldsymbol{\sigma}^2 = (\sigma_e^2, \sigma_v^2)'$ in the case of known λ . Then it is assumed that the estimator $\hat{\boldsymbol{\sigma}}^2(\lambda)$ satisfies the following:

$$(A.3) \quad (\hat{\boldsymbol{\sigma}}^2(\lambda) - \boldsymbol{\sigma}^2)|\mathbf{y}_i = \mathbf{O}_p(m^{-1/2}).$$

$$(A.4) \quad E[\hat{\boldsymbol{\sigma}}^2(\lambda) - \boldsymbol{\sigma}^2|\mathbf{y}_i] = \mathbf{O}_p(m^{-1}).$$

$$(A.5) \quad \partial \hat{\boldsymbol{\sigma}}^2(\lambda) / \partial \lambda | \mathbf{y}_i = \mathbf{O}_p(1).$$

$$(A.6) \quad \left(\partial \hat{\boldsymbol{\sigma}}^2(\lambda) / \partial \lambda - E[\partial \hat{\boldsymbol{\sigma}}^2(\lambda) / \partial \lambda | \mathbf{y}_i] \right) | \mathbf{y}_i = \mathbf{O}_p(m^{-1/2}).$$

Condition (A.3) implies that the estimators $\hat{\sigma}_v^2(\lambda)$ and $\hat{\sigma}_e^2(\lambda)$ are consistent. Conditions (A.4) and (A.8) will be used for investigating asymptotic properties of $\hat{\boldsymbol{\sigma}}^2(\hat{\lambda})$.

Substituting $\hat{\rho}(\lambda) = \hat{\sigma}_v^2(\lambda) / \hat{\sigma}_e^2(\lambda)$ into $\hat{\boldsymbol{\beta}}(\rho, \lambda)$, one gets the estimator $\hat{\boldsymbol{\beta}}(\lambda)$ defined by

$$\hat{\boldsymbol{\beta}}(\lambda) = \hat{\boldsymbol{\beta}}(\hat{\rho}(\lambda), \lambda).$$

It is noted from (A.8) that $\hat{\rho}(\lambda) - \rho = O_p(m^{-1/2})$. Some asymptotic properties on $\hat{\boldsymbol{\beta}}(\lambda)$ are given in the following lemma which will be proved in the Appendix. Lemma 1 will be used in Theorem 1 for showing the second-order accuracy of the parametric bootstrap procedure.

Lemma 1 (Asymptotic properties of $\hat{\boldsymbol{\beta}}(\lambda)$). *Under Assumptions 1 and 2, it holds that $(\hat{\boldsymbol{\beta}}(\lambda) - \boldsymbol{\beta})|\mathbf{y}_i = \mathbf{O}_p(m^{-1/2})$, $E[\hat{\boldsymbol{\beta}}(\lambda) - \boldsymbol{\beta}|\mathbf{y}_i] = \mathbf{O}_p(m^{-1})$ and*

$$\left(\partial \hat{\boldsymbol{\beta}}(\lambda) / \partial \lambda - E[\partial \hat{\boldsymbol{\beta}}(\lambda) / \partial \lambda | \mathbf{y}_i] \right) | \mathbf{y}_i = \mathbf{O}_p(m^{-1/2}).$$

We here provide some conventional estimators of σ_v^2 and σ_e^2 and show whether those estimators satisfy Assumption 2.

[1] **Prasad-Rao type estimator.** Let $\mathbf{X} = (\mathbf{X}'_1, \dots, \mathbf{X}'_m)'$ and $\mathbf{E} = \text{blockdiag}(\mathbf{E}_1, \dots, \mathbf{E}_m)$ for $\mathbf{E}_i = \mathbf{I}_{n_i} - n_i^{-1} \mathbf{J}_i$. Defined $\mathbf{h}(\mathbf{Y}, \lambda)$ by

$$\mathbf{h}(\mathbf{Y}, \lambda) = (\mathbf{h}(\mathbf{y}_1, \lambda)', \dots, \mathbf{h}(\mathbf{y}_m, \lambda)')'.$$

Then define S_1 and S_2 by $S_1 = \mathbf{h}(\mathbf{Y}, \lambda)'(\mathbf{I}_N - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\mathbf{h}(\mathbf{Y}, \lambda)$ and $S_2 = \mathbf{h}(\mathbf{Y}, \lambda)'(\mathbf{E} - \mathbf{E}\mathbf{X}(\mathbf{X}'\mathbf{E}\mathbf{X})^{-1}\mathbf{X}'\mathbf{E})\mathbf{h}(\mathbf{Y}, \lambda)$. Prasad and Rao (1990) suggested unbiased estimators of σ_v^2 and σ_e^2 given by

$$\hat{\sigma}_{e.PR}^2 = \frac{S_2}{N - m - p} \quad \text{and} \quad \hat{\sigma}_{v.PR}^2 = \frac{1}{N_*} \{ S_1 - (N - p)\hat{\sigma}_{e.PR}^2 \}, \quad (8)$$

where $N = \sum_{i=1}^m n_i$ and $N_* = N - \text{tr} \{ (\mathbf{X}'\mathbf{X})^{-1} \sum_{i=1}^m n_i^2 \bar{\mathbf{x}}_i' \bar{\mathbf{x}}_i \}$. The Prasad-Rao estimator of $\boldsymbol{\sigma}^2$ is denoted by $\hat{\boldsymbol{\sigma}}_{PR}^2 = (\hat{\sigma}_{e.PR}^2, \hat{\sigma}_{v.PR}^2)'$. It is noted that $N = O(m)$, $N - m - p = O(m)$ and $N_* = O(m)$ under Assumption 1.

[2] **ML estimator.** The maximum likelihood (ML) estimator $\hat{\boldsymbol{\sigma}}_{ML}^2 = (\hat{\sigma}_{e.ML}^2, \hat{\sigma}_{v.ML}^2)'$ of $\boldsymbol{\sigma}^2 = (\sigma_v^2, \sigma_e^2)'$ are given as the solutions of the equations

$$L_1(\hat{\boldsymbol{\sigma}}_{ML}^2) = 0 \quad \text{and} \quad L_2(\hat{\boldsymbol{\sigma}}_{ML}^2) = 0, \quad (9)$$

where

$$L_1(\boldsymbol{\sigma}^2) = \frac{1}{\sigma_e^4} \sum_{i=1}^m \left\| \mathbf{h}(\mathbf{y}_i, \lambda) - \mathbf{X}_i \hat{\boldsymbol{\beta}}(\rho, \lambda) - \frac{n_i \rho}{1 + n_i \rho} (z_i(\lambda) - \bar{\mathbf{x}}_i' \hat{\boldsymbol{\beta}}(\rho, \lambda)) \mathbf{j}_i \right\|^2 - \sum_{i=1}^m \frac{n_i}{\sigma_e^2} \left(1 - \frac{\rho}{1 + n_i \rho} \right),$$

$$L_2(\boldsymbol{\sigma}^2) = \sum_{i=1}^m \frac{n_i^2}{(\sigma_e^2 + n_i \sigma_v^2)^2} \{ z_i(\lambda) - \bar{\mathbf{x}}_i' \hat{\boldsymbol{\beta}}(\rho, \lambda) \}^2 - \sum_{i=1}^m \frac{n_i}{\sigma_e^2 + n_i \sigma_v^2}.$$

[3] REML estimator. The restricted maximum likelihood (REML) estimator $\hat{\boldsymbol{\sigma}}_{RML}^2 = (\hat{\sigma}_{e.RML}^2, \hat{\sigma}_{v.RML}^2)'$ of $\boldsymbol{\sigma}^2$ is given as the solutions of the equations

$$\begin{aligned} 0 &= L_1(\boldsymbol{\sigma}^2) + \text{tr}[(\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Sigma}^{-2}\mathbf{X}], \\ 0 &= L_2(\boldsymbol{\sigma}^2) + \text{tr}[(\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{Z}\boldsymbol{\Sigma}^{-1}\mathbf{X}], \end{aligned} \quad (10)$$

where $\boldsymbol{\Sigma} = \text{blockdiag}(\sigma_e^2 \mathbf{V}_1(\rho), \dots, \sigma_e^2 \mathbf{V}_m(\rho))$, the covariance matrix of $\mathbf{h}(\mathbf{Y}, \lambda)$.

The following lemma guarantees that the above three estimators satisfy Assumption 2, where the proof will be given in the Appendix.

Lemma 2. *Under Assumption 1, the above three estimators $\hat{\boldsymbol{\sigma}}_{PR}^2$, $\hat{\boldsymbol{\sigma}}_{ML}^2$ and $\hat{\boldsymbol{\sigma}}_{RML}^2$ satisfy Assumption 2.*

It may be guessed from Lemma 2 that Assumption 2 is not so restrictive, because it is satisfied by the three typical estimators.

Finally, we provide an estimator of the transformation parameter λ based on the estimators $\hat{\boldsymbol{\beta}}(\lambda)$, $\hat{\sigma}_v^2(\lambda)$ and $\hat{\sigma}_e^2(\lambda)$. Using the likelihood (6), we suggest the estimator as a solution of the equation

$$F(\hat{\boldsymbol{\beta}}(\lambda), \hat{\sigma}_v^2(\lambda), \hat{\sigma}_e^2(\lambda), \lambda) = 0, \quad (11)$$

where

$$F(\boldsymbol{\beta}, \sigma_v^2, \sigma_e^2, \lambda) = \sigma_e^{-2} \sum_{i=1}^m \left(\mathbf{h}(\mathbf{y}_i, \lambda) - \mathbf{X}_i \boldsymbol{\beta} \right)' \mathbf{V}_i(\rho)^{-1} \left(\frac{\partial}{\partial \lambda} \mathbf{h}(\mathbf{y}_i, \lambda) \right) + J(\mathbf{Y}, \lambda),$$

for

$$J(\mathbf{Y}, \lambda) = \sum_{i=1}^m \sum_{j=1}^{n_i} \frac{h_{x\lambda}(y_{ij}, \lambda)}{h_x(y_{ij}, \lambda)}.$$

Note that $h_{x\lambda}(y_{ij}, \lambda) = \partial^2 h(x, \lambda) / \partial x \partial \lambda|_{x=y_{ij}}$ and $h_x(y_{ij}, \lambda) = \partial h(x, \lambda) / \partial x|_{x=y_{ij}}$.

Lemma 3. *Under Assumptions 1 and 2, the equation (11) includes a solution which is consistent to λ . This solution is denoted by $\hat{\lambda}$. Then, $(\hat{\lambda} - \lambda)|\mathbf{y}_i = O_p(m^{-1/2})$ and $E(\hat{\lambda} - \lambda|\mathbf{y}_i) = O_p(m^{-1})$.*

It is easy to see that $E(\hat{\lambda} - \lambda) = O(m^{-1})$ from Lemma 3 since $E(\hat{\lambda} - \lambda) = E[E(\hat{\lambda} - \lambda|\mathbf{y}_i)]$. Based on the results given in the above lemmas, we can get the following asymptotic properties of estimators for the unknown parameters in TNERM (4). The proof is given in the Appendix.

Lemma 4. *Let $\boldsymbol{\theta} = (\boldsymbol{\beta}', \sigma_v^2, \sigma_e^2, \lambda)'$ and $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\beta}}(\hat{\lambda})', \sigma_v^2(\hat{\lambda}), \sigma_e^2(\hat{\lambda}), \hat{\lambda})'$. Under Assumptions 1 and 2, we have $(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})|\mathbf{y}_i = O_p(m^{-1/2})$ and $E(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}|\mathbf{y}_i) = O_p(m^{-1})$ for $i = 1, \dots, m$.*

The latter property that $E(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}|\mathbf{y}_i) = O_p(m^{-1})$ is technical but crucial for the proof of Theorem 1 in Section 3, which gives validity of the bootstrap method for constructing prediction intervals of TEBLUP.

3 Prediction and its Uncertainty

We now provide the transformed empirical best linear unbiased predictor (TEBLUP) for small area estimation and construct the prediction intervals based on TEBLUP as a measure of uncertainty of the predictor. Since TEBLUP includes the estimators of the parameters β , σ_v^2 , σ_e^2 and λ , it is difficult to construct an exact prediction interval. Thus, in this section, we try to construct a prediction interval with the second-order accuracy. To this end, the asymptotic results given in the lemmas in the previous section are heavily used.

3.1 TEBLUP

We here consider the problem of predicting the quantity

$$h^{-1}(\xi_i, \lambda) \quad \text{for} \quad \xi_i = \bar{\mathbf{x}}_i' \beta + v_i.$$

When $\theta = (\beta, \sigma_v^2, \sigma_e^2, \lambda)$ is known, it is well known that the conditional distribution of ξ_i given \mathbf{y}_i is $\mathcal{N}(\hat{\xi}_i(\theta), \sigma_i^2)$, where

$$\hat{\xi}_i = \hat{\xi}_i(\theta) = E[\xi_i | \mathbf{y}_i] = \bar{\mathbf{x}}_i' \beta + \frac{n_i \rho}{1 + n_i \rho} (z_i(\lambda) - \bar{\mathbf{x}}_i' \beta), \quad (12)$$

and

$$\sigma_i^2 = \sigma_i^2(\sigma^2) = \frac{\sigma_v^2}{1 + n_i \rho}. \quad (13)$$

The estimator $\hat{\xi}_i(\theta)$ is the Bayes estimator of ξ_i in the Bayesian context. Substituting the GLS $\hat{\beta}(\rho, \lambda)$ given in (7) into (12) yields the predictor

$$\bar{\mathbf{x}}_i' \hat{\beta}(\rho, \lambda) + \frac{n_i \rho}{1 + n_i \rho} (z_i(\lambda) - \bar{\mathbf{x}}_i' \hat{\beta}(\rho, \lambda)).$$

It is known that this estimator is the best linear unbiased predictor (BLUP) of ξ_i . For σ_v^2 , σ_e^2 and λ , we substitute the estimators given in Section 2 into the BLUP, and the resulting predictor is given by

$$\hat{\xi}_i^{EB} = \hat{\xi}_i(\hat{\theta}) = \bar{\mathbf{x}}_i' \hat{\beta} + \frac{n_i \hat{\rho}}{1 + n_i \hat{\rho}} (z_i(\hat{\lambda}) - \bar{\mathbf{x}}_i' \hat{\beta}), \quad (14)$$

where, for simplicity, we use the notations $\hat{\beta}$, $\hat{\sigma}_v^2$, $\hat{\sigma}_e^2$ and $\hat{\rho}$ as abbreviation of $\hat{\beta}(\hat{\lambda})$, $\hat{\sigma}_v^2(\hat{\lambda})$, $\hat{\sigma}_e^2(\hat{\lambda})$ $\hat{\rho}(\hat{\lambda}) = \hat{\sigma}_v^2(\hat{\lambda})/\hat{\sigma}_e^2(\hat{\lambda})$ without any confusion. The predictor (14) is called the empirical best linear unbiased predictor (EBLUP). In the Bayesian context, it corresponds to the empirical Bayes estimator of ξ_i .

Since our interest is in the prediction of $h^{-1}(\xi_i, \lambda)$, we need to make the inverse transformation of $\hat{\xi}_i^{EB}$. It should be remarked that the inverse transformation depends on the unknown transformation parameter λ . Hence, the transformed predictor of $h^{-1}(\xi_i, \lambda)$ is given by

$$h^{-1}(\hat{\xi}_i^{EB}, \hat{\lambda}), \quad (15)$$

which is called the transformed empirical best linear unbiased predictor (TEBLUP).

3.2 Prediction interval based on TEBLUP

For measuring uncertainty of TEBLUP, we propose prediction intervals of $h^{-1}(\xi_i, \lambda)$ with a second-order accuracy for $\xi_i = \mathbf{x}'_i \boldsymbol{\beta} + v_i$. The basic idea of constructing prediction intervals are based on Chatterjee, et al. (2008), who proposed the parametric bootstrap method for constructing a second-order accurate unconditional confidence interval in normal linear mixed models.

Recall that conditionally $\xi_i | \mathbf{y}_i \sim \mathcal{N}(\hat{\xi}_i(\boldsymbol{\theta}), \sigma_i^2)$, where $\hat{\xi}_i(\boldsymbol{\theta})$ and σ_i^2 are given in (12) and (13), respectively. The conditional distribution given \mathbf{y}_i implies that

$$\sigma_i^{-1} \{h(h^{-1}(\xi_i, \lambda), \lambda) - \hat{\xi}_i(\boldsymbol{\theta})\}$$

is a standard normal pivot since $h(h^{-1}(\xi_i, \lambda), \lambda) = \xi_i$.

Let $\hat{\sigma}_i^2 = \hat{\sigma}_v^2 / (1 + n_i \hat{\rho})$. For $\hat{\xi}_i^{EB}$ given in (14), we want to obtain a distribution of

$$T_i = T_i(\xi_i, \lambda, \hat{\boldsymbol{\theta}}, \mathbf{y}_i) = \hat{\sigma}_i^{-1} \{h(h^{-1}(\xi_i, \lambda), \hat{\lambda}) - \hat{\xi}_i^{EB}\}. \quad (16)$$

This distribution is denoted by \mathcal{L}_m . If there were constants a_α and b_α such that $P[a_\alpha \leq \hat{\sigma}_i^{-1} \{h(h^{-1}(\xi_i, \lambda), \hat{\lambda}) - \hat{\xi}_i^{EB}\} \leq b_\alpha] = 1 - \alpha$, one would get a $100(1 - \alpha)\%$ prediction interval

$$h^{-1}(\xi_i, \lambda) \in [h^{-1}(\hat{\xi}_i^{EB} + a_\alpha \hat{\sigma}_i, \hat{\lambda}), h^{-1}(\hat{\xi}_i^{EB} + b_\alpha \hat{\sigma}_i, \hat{\lambda})].$$

However, $h(h^{-1}(\xi_i, \lambda), \hat{\lambda})$ is directly affected by the randomness of $\hat{\lambda}$, and the distribution \mathcal{L}_m of (16) depends on unknown parameters. Thus, a_α and b_α are not free from unknown parameters. A feasible approach is an asymptotic approximation of \mathcal{L}_m . Since the estimator $\hat{\boldsymbol{\theta}}$ is consistent from Lemma 4, it can be seen that \mathcal{L}_m converges to the standard normal distribution as m tends to infinity. By approximating a_α and b_α with quantiles of the standard normal distribution, we can construct a prediction interval of $h^{-1}(\xi_i, \lambda)$. However, the accuracy of this prediction interval can be confirmed that order $O(m^{-1})$, so that such an approximation does not guarantee enough accuracy.

To obtain a prediction interval with accuracy up to $O(m^{-3/2})$, we consider to estimate the distribution \mathcal{L}_m based on the parametric bootstrap method. Let y_{ij}^* 's be a bootstrap sample which is generated as

$$y_{ij}^* = h^{-1}(\mathbf{x}'_{ij} \hat{\boldsymbol{\beta}} + v_i^* + \varepsilon_{ij}^*, \hat{\lambda}), \quad i = 1, \dots, m, \quad j = 1, \dots, n_i,$$

where v_i^* 's and ε_{ij}^* 's are mutually independently distributed as $v_i^* \sim \mathcal{N}(0, \hat{\sigma}_v^2)$ and $\varepsilon_{ij}^* \sim \mathcal{N}(0, \hat{\sigma}_e^2)$. The estimator $\hat{\boldsymbol{\theta}}^* = ((\hat{\boldsymbol{\beta}}^*)', \hat{\sigma}_v^{2*}, \hat{\sigma}_e^{2*}, \hat{\lambda}^*)'$ is calculated from y_{ij}^* 's with the same methods as used to obtain $\hat{\boldsymbol{\theta}}$. Let $\hat{\xi}_i^{EB*} = \bar{\mathbf{x}}'_i \hat{\boldsymbol{\beta}}^* + (n_i \hat{\rho}^* / (1 + n_i \hat{\rho}^*)) (z_i^*(\hat{\lambda}^*) - \bar{\mathbf{x}}'_i \hat{\boldsymbol{\beta}}^*)$ and $\hat{\sigma}_i^{2*} = \hat{\sigma}_v^{2*} / (1 + n_i \hat{\rho}^*)$ for $\hat{\rho}^* = \hat{\sigma}_v^{2*} / \hat{\sigma}_e^{2*}$ and $z_i^*(\hat{\lambda}^*) = n_i^{-1} \sum_{j=1}^{n_i} h(y_{ij}^*, \hat{\lambda}^*)$. For $\xi_i^* = \bar{\mathbf{x}}'_i \hat{\boldsymbol{\beta}}^* + v_i^*$, consider the distribution of

$$T_i^* = (\hat{\sigma}_i^*)^{-1} \{h(h^{-1}(\xi_i^*, \hat{\lambda}^*), \hat{\lambda}^*) - \hat{\xi}_i^{EB*}\}, \quad (17)$$

which is denoted by \mathcal{L}_m^* . As shown in Theorem 1 given below, the distribution \mathcal{L}_m in (16) can be approximated by the bootstrap distribution \mathcal{L}_m^* with accuracy of order $O_p(m^{-3/2})$. Using this approximation, we then proceed to obtain a prediction interval.

Theorem 1. *Under Assumptions 1 and 2, we have*

$$\sup_{q \in \mathbb{R}} |\mathcal{L}_m(q) - \mathcal{L}_m^*(q)| = O_p(m^{-3/2}). \quad (18)$$

The proof of Theorem 1 is given in the Appendix. A direct application of Theorem 1 is the following result on highly accurate prediction intervals.

Corollary 1. *For any $\alpha \in (0, 1)$, let $q_1 = q_1(\mathbf{Y})$ and $q_2 = q_2(\mathbf{Y})$ be appropriate quantiles based on the bootstrap sample such that*

$$\mathcal{L}_m^*(q_2) - \mathcal{L}_m^*(q_1) = 1 - \alpha,$$

where $\mathcal{L}_m^*(\cdot)$ is the distribution function of T_i^* . Then, one gets the prediction interval of $h^{-1}(\xi_i, \lambda)$ given by

$$I_m = [h^{-1}(\hat{\xi}_i^{EB} + q_1 \hat{\sigma}_i, \hat{\lambda}), h^{-1}(\hat{\xi}_i^{EB} + q_2 \hat{\sigma}_i, \hat{\lambda})]. \quad (19)$$

Under Assumptions 1 and 2, it holds that

$$P(h^{-1}(\xi_i, \lambda) \in I_m) = 1 - \alpha + O(m^{-3/2}). \quad (20)$$

Corollary 1 gives us a highly accurate prediction interval of $h^{-1}(\xi_i, \lambda)$ based on TEBLUP. The prediction interval I_m implies that one can figure out precision of TEBLUP with the length of the interval I_m . It is also noted that the coverage accuracy of the prediction interval given in Corollary 1 can be further improved up to $O(m^{-5/2})$ with one round of calibration.

3.3 Conditional prediction interval

We next construct a conditional prediction interval given data in the area of interest. When data \mathbf{y}_i are observed from the i -th area, Booth and Hobert (1998), Datta, Kubokawa, Molina and Rao (2011), Sugawara and Kubokawa (2014) and Torabi and Rao (2013) treated the conditional MSE of the EBLUP $\hat{\xi}_i^{EB}$ given \mathbf{y}_i , namely, $E[(\hat{\xi}_i^{EB} - \xi_i)^2 | \mathbf{y}_i]$. This conditional MSE measures how much the EBLUP has an estimation error given the data \mathbf{y}_i , and this conditional approach may be appealing because it conditions on the data in the area of interest. In this subsection, we construct a conditional prediction interval I_m^c given \mathbf{y}_i such that

$$P(h^{-1}(\xi_i, \lambda) \in I_m^c | \mathbf{y}_i) = 1 - \alpha + O_p(m^{-3/2}). \quad (21)$$

To this end, we need to approximate the conditional distribution of $T_i = T_i(\xi_i, \lambda, \hat{\theta}, \mathbf{y}_i) = \hat{\sigma}_i^{-1} \{h(h^{-1}(\xi_i, \lambda), \hat{\lambda}) - \hat{\xi}_i^{EB}\}$ given \mathbf{y}_i . Denote this conditional distribution by $\mathcal{L}_m^c = \mathcal{L}_m^c(\cdot | \mathbf{y}_i)$. The difference between the unconditional and conditional prediction intervals is that the unconditional distribution of T_i is considered in (16), while the conditional distribution of T_i given \mathbf{y}_i is treated. It is noted that there is a correlation between ξ_i and \mathbf{y}_i in (21), namely, the conditional distribution of ξ_i given \mathbf{y}_i is $\mathcal{N}(\hat{\xi}_i(\boldsymbol{\theta}), \sigma_i^2)$ for $\hat{\xi}_i(\boldsymbol{\theta})$ and σ_i^2 given in (12) and (13).

Since it is difficult to derive an exact conditional distribution of T_i given \mathbf{y}_i , we suggest to approximate it via the parametric bootstrap method. A bootstrap sample is generated as

$$y_{kj}^* = h^{-1}(\mathbf{x}_{kj}' \hat{\boldsymbol{\beta}} + v_k^* + \varepsilon_{kj}^*, \hat{\lambda}), \quad k \neq i, k = 1, \dots, m, \quad j = 1, \dots, n_k,$$

where v_k^* 's and ε_{kj}^* 's are mutually independently distributed as $v_k^* \sim \mathcal{N}(0, \hat{\sigma}_v^2)$ and $\varepsilon_{kj}^* \sim \mathcal{N}(0, \hat{\sigma}_e^2)$. Let $\mathbf{y}_k^* = (y_{k1}^*, \dots, y_{kn_k}^*)'$ for $k \neq i$. Noting that \mathbf{y}_i is fixed, we can construct the estimator $\hat{\boldsymbol{\theta}}_{(i)}^* = ((\hat{\boldsymbol{\beta}}_{(i)}^*)', \hat{\sigma}_{v(i)}^{2*}, \hat{\sigma}_{e(i)}^{2*}, \hat{\lambda}_{(i)}^*)'$ from

$$\mathbf{y}_1^*, \dots, \mathbf{y}_{i-1}^*, \mathbf{y}_i, \mathbf{y}_{i+1}^*, \dots, \mathbf{y}_m^*,$$

with the same technique as used to obtain $\hat{\boldsymbol{\theta}}$. Let $\hat{\xi}_{(i)}^{EBc*} = \bar{\mathbf{x}}_i' \hat{\boldsymbol{\beta}}_{(i)}^* + (n_i \hat{\rho}_{(i)}^* / (1 + n_i \hat{\rho}_{(i)}^*)) (z_i(\hat{\lambda}_{(i)}^*) - \bar{\mathbf{x}}_i' \hat{\boldsymbol{\beta}}_{(i)}^*)$ and $\hat{\sigma}_{(i)}^{2c*} = \hat{\sigma}_{v(i)}^{2*} / (1 + n_i \hat{\rho}_{(i)}^*)$ for $\hat{\rho}_{(i)}^* = \hat{\sigma}_{v(i)}^{2*} / \hat{\sigma}_{e(i)}^{2*}$ and $z_i(\hat{\lambda}_{(i)}^*) = n_i^{-1} \sum_{j=1}^{n_i} h(y_{ij}, \hat{\lambda}_{(i)}^*)$. Let ξ_i^{c*} be a random variable having $\mathcal{N}(\hat{\xi}_i^{EB}, \hat{\sigma}_i^2)$ for $\hat{\xi}_i^{EB} = \hat{\xi}_i(\hat{\boldsymbol{\theta}})$. Then, for fixed \mathbf{y}_i , we consider the distribution of

$$T_{(i)}^{c*} = (\hat{\sigma}_{(i)}^{c*})^{-1} \{h(h^{-1}(\xi_i^{c*}, \hat{\lambda}), \hat{\lambda}_{(i)}^*) - \hat{\xi}_{(i)}^{EBc*}\}, \quad (22)$$

which is denoted by $\mathcal{L}_m^{c*} = \mathcal{L}_m^{c*}(\cdot | \mathbf{y}_i)$. Similarly to the unconditional case, we can obtain a conditional prediction interval via the parametric bootstrap approximation.

Theorem 2. *Under Assumptions 1 and 2, we have*

$$\sup_{q \in \mathbb{R}} |\mathcal{L}_m^c(q | \mathbf{y}_i) - \mathcal{L}_m^{c*}(q | \mathbf{y}_i)| = O_p(m^{-3/2}). \quad (23)$$

The proof of Theorem 2 is given in the Appendix. From Theorem 2, we obtain a conditional prediction interval with second-order accuracy.

Corollary 2. *For any $\alpha \in (0, 1)$, let $q_1^c = q_1^c(\mathbf{Y})$ and $q_2^c = q_2^c(\mathbf{Y})$ be appropriate quantiles based on the bootstrap sample such that*

$$\mathcal{L}_m^{c*}(q_2^c | \mathbf{y}_i) - \mathcal{L}_m^{c*}(q_1^c | \mathbf{y}_i) = 1 - \alpha,$$

where $\mathcal{L}_m^{c*}(\cdot | \mathbf{y}_i)$ is the distribution function of $T_{(i)}^{c*}$. Then, one gets the prediction interval of $h^{-1}(\xi_i, \lambda)$ given by

$$I_m^c = [h^{-1}(\hat{\xi}_i^{EB} + q_1^c \hat{\sigma}_i, \hat{\lambda}), h^{-1}(\hat{\xi}_i^{EB} + q_2^c \hat{\sigma}_i, \hat{\lambda})]. \quad (24)$$

Under Assumptions 1 and 2, it holds that

$$P(h^{-1}(\xi_i, \lambda) \in I_m^c | \mathbf{y}_i) = 1 - \alpha + O_p(m^{-3/2}). \quad (25)$$

4 Simulation Studies

In this section, we investigate performances of the procedures suggested in the previous sections by Monte Carlo simulation.

4.1 Disadvantages under misspecified transformation

We investigate disadvantages when we use a misspecified transformation under DPT or DPLT.

[TNERM with DPT] Since the purpose of small area estimation is to predict the mean of each small area as accurately as possible, we begin by comparing the prediction error among

TNERM with DPT, the log-transformed NERM and the non-transformed NERM. We generate observations from the model $h(y_{ij}, \lambda) = \beta_0 + \beta_1 x_{ij} + v_i + \varepsilon_{ij}$ for $i = 1, \dots, 20$ and $j = 1, 2, 3$, where v_i and ε_{ij} are generated from $\mathcal{N}(0, (0.5)^2)$ with $\sigma_v^2 = 0.5$ and $\mathcal{N}(0, 1)$ with $\sigma_e^2 = 1$, respectively, x_{ij} are generated from $\mathcal{N}(0, 1)$, which are fixed through simulation runs, and $\beta_0 = \beta_1 = 1$. We consider the 6 patterns of λ . We note that the log-transformation is a misspecified transformation when $\lambda \neq 0$, and the identity transformation is always misspecified regardless of λ . The true values we want to predict is

$$h^{-1}(\beta_0 + \beta_1 \bar{x}_i + v_i, \lambda).$$

Let $\hat{\theta}_i^{(1)}$ be the TEBLUP defined in (15) and let $\hat{\theta}_i^{(2)}$ and $\hat{\theta}_i^{(3)}$ be predictors based on the log-transformed NERM and the non-transformed NERM, respectively. Then we can define the prediction mean squared error (PMSE) as

$$\text{PMSE}_k = \sum_{i=1}^m E \left[\left\{ \hat{\theta}_i^{(k)} - h^{-1}(\beta_0 + \beta_1 \bar{x}_i + v_i, \lambda) \right\}^2 \right], \quad k = 1, 2, 3$$

which can be estimated based on $R = 2000$ simulation runs. To see the differences among PMSE_k , we calculate the improvement ratio of PMSE (IRP) defined as

$$\text{IRP}_k = \frac{\text{PMSE}_1 - \text{PMSE}_k}{\text{PMSE}_k} \times 100, \quad k = 2, 3.$$

The simulation results are given in Table 1. It is observed that PMSE in log-NERM gets worse than that in TNERM when λ is away from 0. When $\lambda = 0$, the log-transformation is the true transformation, so that it is natural that PMSE in TNERM is slightly larger than that in log-NERM since there is an estimation error of λ in TNERM. In the conventional non-transformed NERM, the prediction errors are always bad compared to the other two models, but gets better as λ gets larger. This is because the DPT is similar to the identity transformation when λ is close to 1.

We next investigated the percentage of zero estimates of σ_v^2 , variance of random effects. Zero estimates of σ_v^2 means that the resulting EBLUP estimator of a small area mean is over-shrunk to the regression estimator. Such a situation is not preferable in practice and recently Li and Lahiri (2010) and Yoshimori and Lahiri (2014) discussed this problem. Based on the same simulation runs, we calculate the percentage of zero estimates of σ_v^2 . We report the simulation result in Table 1. The percentage in log-NERM is slightly larger compared to TNERM, and the percentage in conventional NERM are always large but gets smaller as λ goes to 1, which seems to be the same reason in the prediction error.

[TNERM with DPLT] We next consider to compare the performances among the proposed TNERM with DPLT, the logistic-transformed NERM and the non-transformed NERM. We generate observations from the model $h(y_{ij}, \lambda) = \beta_0 + \beta_1 x_{ij} + v_i + \varepsilon_{ij}$, where v_i and ε_{ij} are generated from $\mathcal{N}(0, 0.2^2)$ and $\mathcal{N}(0, 0.3^2)$, respectively, and x_{ij} are generated from $\mathcal{N}(0, 1)$, which are fixed through simulation runs, and $\beta_0 = -1$ and $\beta_1 = 1$. We consider the 6 patterns of λ as same in the previous study. For each λ , we similarly compute IRP_k , $k = 2, 3$ and the percentage of zero estimates of σ_e^2 based on $K = 2,000$ simulation runs, where the results are provided in Table 2.

Table 1: The Percentage Ratio of Prediction Errors in TNERM with DPT Compared to Log-transformed NERM and NERM, and the Percentage of Zero Estimates of σ_v^2 for Three Models.

λ	improvement of PMSE		percentage of zero estimates of σ_v^2		
	IRP ₂	IRP ₃	TNERM	log-NERM	NERM
0.0	+1.0	-86.0	2.8	2.8	20.4
0.2	-0.4	-74.5	2.2	2.3	14.0
0.4	-5.1	-54.0	2.0	2.2	7.5
0.6	-10.8	-35.8	2.1	2.5	5.4
0.8	-15.1	-20.3	2.9	3.8	3.9
1.0	-18.1	-10.3	1.9	3.1	3.0

From Table 2, we can observe that the prediction error in logistic-NERM is worse than that in TNERM when λ is far from 0. We also note that IRP₂ in the DPLT case decreases more rapidly than that in the DPT case. Moreover, it is important to point out that the prediction errors when λ is close to 0 are almost the same, so that we do not have much disadvantage in using TNERM in terms of prediction errors. For the non-transformed NERM, the prediction errors are always poor compared to other two models. It is pointed out that the estimating rate of $\hat{\sigma}_e^2 = 0$ are small, but the rates in logistic-NERM and the conventional NERM tends to be larger than that of TNERM.

Table 2: The Percentage Ratio of Prediction Errors in TNERM with DPLT Compared to Logistic-transformed NERM and NERM, and the Percentage of Zero Estimates of σ_v^2 for Three Models.

λ	improvement of PMSE		percentage of zero estimates of σ_v^2		
	IRP ₂	IRP ₃	TNERM	logistic-NERM	NERM
0.0	+2.8	-67.8	0.6	0.6	1.2
0.2	-8.6	-67.5	0.7	1.3	1.3
0.4	-29.4	-67.8	0.9	0.8	1.0
0.6	-42.4	-68.8	1.1	1.1	1.6
0.8	-50.2	-69.4	0.8	1.2	1.6
1.0	-54.7	-69.4	0.9	1.4	1.9

4.2 Finite sample behavior of proposed prediction intervals

In this section, we investigate finite performances of the unconditional and conditional prediction intervals suggested in the previous section for the DP and DPL transformations. The performances are examined by Monte Carlo simulation in the case of $\mathbf{x}'_{ij}\boldsymbol{\beta} = \mu$ without covariates as treated in Chatterjee, etal. (2008).

In the simulation experiments, 1,000 observations for y_{ij} are generated as $y_{ij} = h^{-1}(v_i + \varepsilon_{ij}, \lambda)$, $i = 1, \dots, m$, $j = 1, \dots, n_i$, for $m = 10$, $n_i = 5$ and $\lambda = 0, 0.5$ and 1.0 , where v_i 's and ε_{ij} 's are mutually independently generated from $\mathcal{N}(0, 1)$ with $\mu = 0$, $\sigma_v^2 = 1$ and $\sigma_e^2 = 1$. The frequency of the prediction interval which includes $h^{-1}(\xi_i, \lambda)$ is counted for $i = 1, \dots, m$, and the coverage probability is estimated by dividing the total number of the frequency by 1,000, where the size of the bootstrap sample is 200. The expected length of the prediction interval can be also estimated as an average length by a similar method.

Under the above simulation, we investigate the performances of the unconditional prediction interval and compare it with the naive prediction interval given by

$$[h^{-1}(\hat{\xi}_i^{EB} - z_{\alpha/2}\hat{\sigma}_i, \hat{\lambda}), h^{-1}(\hat{\xi}_i^{EB} + z_{\alpha/2}\hat{\sigma}_i, \hat{\lambda})], \quad (26)$$

where $z_{\alpha/2}$ is the upper $\alpha/2$ quantile point of the standard normal distribution. This is an empirical Bayes confidence interval which is derived by substituting the estimators into the Bayes confidence interval. The maximum likelihood estimators are used for the variance components σ_v^2 and σ_e^2 . Table 3 reports the coverage probability (CP) and the expected length (EL) of the two unconditional prediction intervals (19) and (26) based on the bootstrap method (BT) and the naive method (NV). From Table 3, it is observed that the naive prediction interval is not appropriate since it does not satisfy the nominal confidence coefficient $1 - \alpha = 0.95$, while it gives a shorter length than BT. The prediction interval (19) based on BT has the coverage probability close to the nominal level 0.95. This shows that the correction by the bootstrap method works well.

Table 3: Values of Coverage Probability and Expected Length of the Unconditional Prediction Interval with Confidence Coefficient $1 - \alpha = 0.95$

		λ		0		0.5		1	
				NV	BT	NV	BT	NV	BT
DPT	CP			91.7	96.1	90.6	94.2	90.6	96.3
	EL			2.64	3.91	2.00	2.40	1.50	5.50
DPLT	CP			91.0	95.4	91.2	95.6	90.7	95.4
	EL			0.32	0.41	0.30	0.45	0.27	0.37

We next investigate a performance of the conditional prediction interval given in (24). The same simulation setup as used above is treated for $\lambda = 0.5$ except for fixing initial values of \mathbf{y}_i 's. We first generate initial observations of \mathbf{y}_i 's from the model described above for $i = 1, \dots, 10$ and fix them. Then, the conditional prediction intervals given \mathbf{y}_i are constructed based on the quantiles of the parametric bootstrap samples. The coverage probability (CP) and the expected length (EL) of the conditional prediction interval are reported in Table 4 for TNERM with DPT and DPLT, where the values of \bar{y}_i are the averages \bar{y}_i of the given values \mathbf{y}_i for 10 areas. From Table 4, it is revealed that the coverage probabilities of the conditional prediction intervals are close to the nominal level 0.95 for DPT and DPLT. It is interesting to point out

that the expected length of the conditional prediction interval for DPT is larger as the value of \bar{y}_i is larger, while the expected length in the case of DPLT is not affected by the value of \bar{y}_i . This property of the conditional prediction interval for DPLT is quite different from the unconditional prediction interval.

Table 4: Values of Coverage Probability and Expected Length of the Conditional Prediction Interval with Confidence Coefficient $1 - \alpha = 0.95$

	area	1	2	3	4	5	6	7	8	9	10
	\bar{y}_i	0.44	0.45	1.74	1.91	2.11	2.25	2.47	3.46	3.84	6.58
DPT	CP	95.1	94.9	97.0	95.4	95.5	96.3	97.8	96.2	95.2	93.4
	EL	0.90	0.84	3.33	3.48	3.79	3.95	4.30	5.33	5.23	6.12
	\bar{y}_i	0.26	0.38	0.44	0.47	0.49	0.50	0.52	0.60	0.66	0.68
DPLT	CP	95.2	95.7	94.9	96.3	95.5	97.4	95.1	97.7	96.6	96.2
	EL	0.33	0.41	0.42	0.47	0.41	0.59	0.43	0.49	0.39	0.37

5 Empirical Studies

5.1 Application to the survey data in Japan

We apply the proposed TNERM with DPT to the data in the Survey of Family Income and Expenditure (SFIE) in Japan. In this study, we use the data of the spending item 'Education' in the survey in November 2009, 2010 and 2011 and we are interested in the mean of survey in November. The average spending (scaled by 10,000 Yen) at each capital city of 47 prefectures in Japan is obtained by y_{ij} for $j = 1, 2, 3$ and $i = 1, \dots, 47$. Although the average spendings in SFIE are reported every month, the sample sizes are around 100 for most prefectures, and data of the item 'Education' have high variability. On the other hand, we have data in the National Survey of Family Income and Expenditure (NSFIE) for 47 prefectures. Since NSFIE is based on much larger sample than SFIE, the average spendings in NSFIE are more reliable, but this survey has been implemented every five years. In this study, we use the data of the item 'Education' of NSFIE in 2009, which is denoted by x_i for $i = 1, \dots, 47$. Thus, we apply the TNERM with DPT, that is

$$\frac{y_{ij}^\lambda - y_{ij}^{-\lambda}}{2\lambda} = \beta_0 + \beta_1 \log(x_i) + v_i + \varepsilon_{ij}, \quad j = 1, 2, 3, \quad i = 1, \dots, 47.$$

We used the maximum likelihood estimators for estimation of σ_e and σ_v , and their estimates are $\hat{\sigma}_e^2 = 0.38$ and $\hat{\sigma}_v^2 = 0.13$. The GLS estimates of β_0 and β_1 are $\hat{\beta}_0 = 0.07$ and $\hat{\beta}_1 = -0.001$. The values of EBLUP in ten selected prefectures are reported in Table 5 with the obtained prediction intervals based on proposed procedure given in Sections 3.2 and 3.3 based on 1,000

bootstrap samples. It is observed that the lengths of the unconditional prediction intervals are close each other because of the same sample size. On the other hand, the lengths of the conditional prediction intervals have larger variability than unconditional ones since the conditional prediction intervals depend on not only sample sizes but also the observed values. Note that the estimate of λ is 0.67, which is far away from 0. This means that the logarithmic transformation does not seem appropriate for analyzing the data treated here and we might make predictions inappropriately if we use the logarithmic transformation or we use the original data as discussed in section 4.1.

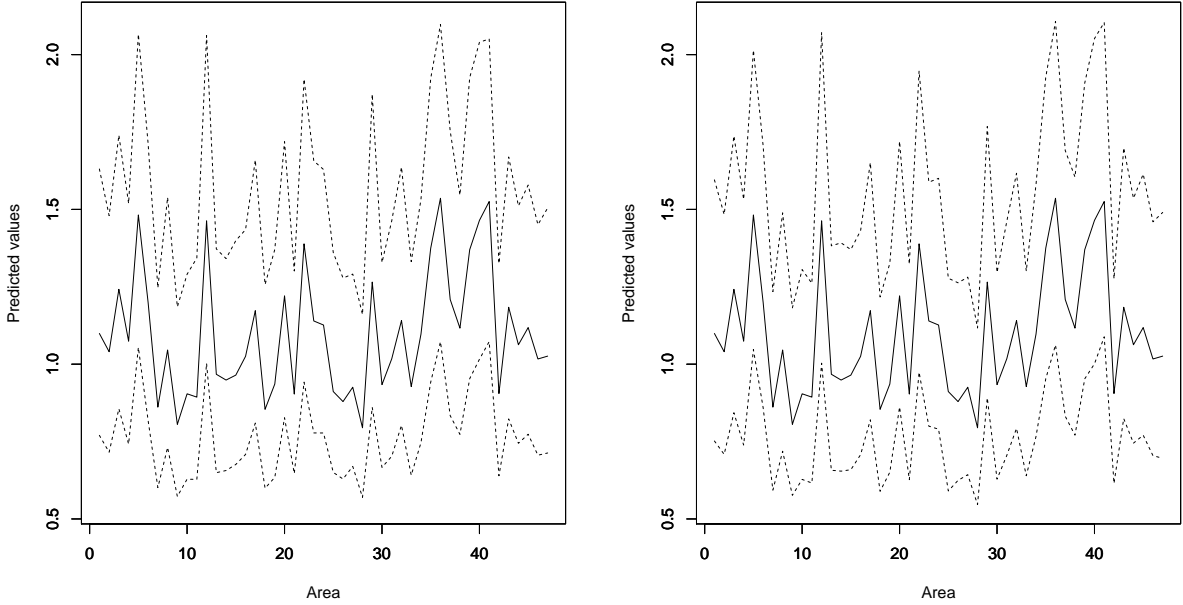


Figure 1: Conditional (right) and Unconditional (left) Prediction Intervals in TNERM with DPT for Survey Data (The solid line denotes TEBLUP and the two dotted lines denote the upper and lower bounds of prediction intervals of each county.)

5.2 Application to the crop areas data

We next deal with the crop areas data with $m = 12$ given by Battese, et al. (1988), which have been used repeatedly in the literature. From the i -th county, n_i segments are sampled for $i = 1, \dots, 12$ and each segment is about 250 hectares, and the area of corn (or soybeans) in the j -th segment, denoted by y_{ij} , is reported as survey data by interviewing farm operators. For the j -th segment, on the other hand, the numbers of pixels (0.45 hectares) classified as corn and soybeans, denoted by x_{1ij} and x_{2ij} , are available from LANDSAT satellite data. Battese, et al. (1988) analyzed the data successfully using the nested error regression model (NERM) in the framework of a finite population, but in the analysis here, we do not assume the finite population model. It is clear from characteristics of the data that y_{ij} 's are bounded above from 250 hectares, which means that the scaled observation $z_{ij} = y_{ij}/250$ lies in the interval $(0, 1)$,

Table 5: Values of TEBLUP for DPT and the Length of Prediction Intervals for Selected Ten Areas.

Area	sample size	sample mean	TEBLUP	lengths of intervals	
				CI	cCI
2	3	1.10	1.06	1.07	1.16
9	3	0.72	0.96	1.01	0.75
12	3	1.78	1.20	1.50	1.57
16	3	1.09	1.05	1.14	1.08
22	3	1.61	1.19	1.27	1.44
28	3	0.70	0.95	1.01	0.87
29	3	1.54	1.14	1.52	1.66
35	3	1.56	1.19	1.34	1.22
39	3	1.53	1.19	1.33	1.42
47	3	1.01	1.06	1.18	1.10

and we apply TNERM with DPLT, described as

$$(2\lambda)^{-1} \left\{ \left(\frac{z_{ij}}{1 - z_{ij}} \right)^\lambda - \left(\frac{z_{ij}}{1 - z_{ij}} \right)^{-\lambda} \right\} = \beta_0 + \beta_1 \log \left(\frac{x_{1ij}^*}{1 - x_{1ij}^*} \right) + \beta_2 \log \left(\frac{x_{2ij}^*}{1 - x_{2ij}^*} \right) + v_i + \varepsilon_{ij}, \quad (27)$$

where $x_{1ij}^* = 0.45x_{1ij}/250$ and $x_{2ij}^* = 0.45x_{2ij}/250$. Note that $0.45x_{1ij}$ and $0.45x_{2ij}$ are bounded above from 250. The estimates of the parameters by the maximum likelihood method are given by $\hat{\beta} = (-0.23, 0.85, -0.05)'$, $\hat{\sigma}_v = 0.11$, $\hat{\sigma}_e = 0.28$ and $\hat{\lambda} = 0.37$. Since the estimate of λ is away from 0, the standard logistic transformation is not appropriate in the framework of model (27). It is noted that the estimate $\hat{\beta}_2$ is close to zero, which implies that the survey data of corn areas are not affected by x_{2ij} , the number of pixels of soybeans. Based on 1,000 bootstrap samples, we get the unconditional and conditional prediction intervals, which are illustrated in Figure 5.2. Figure 5.2 and Table 5.2 show that the length of unconditional prediction intervals are shorter as the sample size is larger, but the conditional prediction intervals do not have the similar properties because the conditional prediction intervals depend on the values of observations.

6 Concluding Remarks

In this paper, we have suggested the parametric transformed nested error regression model (TNERM) as a new unit-level model for analysis of positive or bounded small area data. We have provided the procedures for estimating unknown parameters including the transformation parameter as well as regression coefficients and the variance components. As parametric transformations, we consider the dual power transformation for positive data and the dual power logistic transformation, which we newly proposed in this paper, for bounded data. The transformed EBLUP (TEBLUP) has been made based on the consistent estimators, and unconditional and conditional prediction intervals with second-order accuracy have been constructed

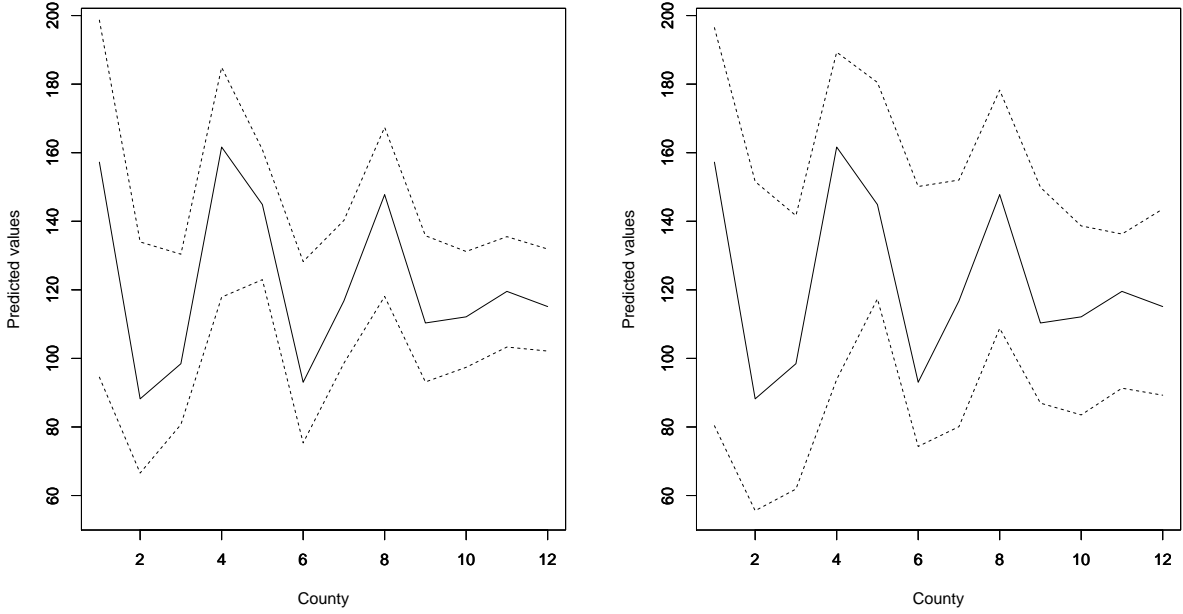


Figure 2: Conditional (right) and Unconditional (left) Prediction Intervals in TNERM with DPLT for Crop Areas Data (The solid line denotes TEBLUP and the two dotted lines denote the upper and lower bounds of prediction intervals of each county.)

based on the parametric bootstrap method. Through the simulation, it is confirmed that prediction accuracy can be bad and the percentage of zero estimation of σ_v^2 tends to be large under misspecified transformation, which motivated us to take the transformation parameter into account of model formulation. The finite sample performances of proposed prediction intervals have been confirmed by simulation as well and the coverage probability of the suggested prediction intervals is close to the nominal level 0.95. For real data applications, we applied TNERM to survey (positive) data in Japan and famous crop areas (bounded) data treated in Battese, et al. (1988). The crucial properties of DPT and DPLT for giving validity of proposed methodology in this paper are summarized in the Appendix and we can use another parametric transformation as alternative to DPT or DPLT whenever it holds the properties.

Our proposed methodology based on the parametric transformation is regarded as a new framework to cope with small-area data, and we hope further development will be studied from theoretical and practical aspects in statistical inferences.

Appendix

All the lemmas and theorems given in this paper will be proved here.

For notational convenience, let $h_{a_1 a_2 \dots a_n}(x, \lambda)$ for $a_1, a_2, \dots, a_n \in \{x, \lambda\}$ be the partial derivative of $h(x, \lambda)$, i.e. $h_{a_1 a_2 \dots a_n}(x, \lambda) = \partial^n f(x, \lambda) / \partial a_1 \dots \partial a_n$. For example, $h_{xx}(x, \lambda) = \partial^2 h(x, \lambda) / \partial x^2$, $h_{x\lambda}(x, \lambda) = \partial^2 h(x, \lambda) / \partial x \partial \lambda$ and others. Moreover $h_{a_1 a_2 \dots a_n}(c, \lambda)$ or $h_{a_1 a_2 \dots a_n}(x, c)$ for $c \in \mathbb{R}$

Table 6: Predicted Hectares of Corn via TEBLUP for DPLT and the Length of Prediction Intervals

County	sample size	sample mean	TEBLUP	lengths of intervals	
				CI	cCI
Cerro Gordo	1	165.8	157.2	111.4	116.7
Hamilton	1	96.3	88.2	66.2	92.7
Worth	1	76.1	98.5	47.9	85.9
Humboldt	2	150.9	161.6	64.8	90.1
Franklin	3	158.6	144.9	37.2	52.3
Pocahontas	3	102.5	93.0	52.5	70.4
Winnebago	3	112.8	116.8	42.8	64.4
Wright	3	144.3	147.8	47.3	61.2
Webster	4	117.6	110.3	43.2	67.2
Hancock	5	109.4	112.1	34.3	55.0
Kossuth	5	110.3	119.6	32.0	39.7
Hardin	5	114.8	115.1	31.4	61.3

means that $h_{a_1 a_2 \dots a_n}(x, \lambda)|_{x=c}$ or $h_{a_1 a_2 \dots a_n}(x, \lambda)|_{\lambda=c}$ respectively.

In their proofs, the following fact will be heavily used: Assume that for $i = 1, \dots, m$, a function $\psi(\mathbf{y}_i)$ is independent of \mathbf{y}_j for $j \neq i$, and that $\psi(\mathbf{y}_i) = O_p(1)$ and $E[\psi(\mathbf{y}_i)] = O(1)$. Then, it follows from the Law of Large Numbers (LLN) that

$$\frac{1}{m} \sum_{j=1}^m \psi(\mathbf{y}_j) \Big| \mathbf{y}_i = O_p(1), \quad i = 1, \dots, m. \quad (28)$$

Moreover, if $E[\psi^2(\mathbf{y}_i)] = O(1)$, then from the Central Limit Theorem (CLT), one gets

$$\frac{1}{\sqrt{m}} \left(\sum_{j=1}^m \psi(\mathbf{y}_j) - E \left[\sum_{j=1}^m \psi(\mathbf{y}_j) \Big| \mathbf{y}_i \right] \right) \Big| \mathbf{y}_i = O_p(1), \quad i = 1, \dots, m, \quad (29)$$

where $(\cdot)|\mathbf{y}_i$ denotes a random variable given \mathbf{y}_i . In the proofs, for notational simplicity, we treat the case of $i = m$ without loss of generality.

Furthermore, we use the properties of DPT and DPLT described in the following.

[Properties of DPT and DPLT] Let the transformation function $h(x, \lambda)$ be $h^{\text{DP}}(x, \lambda)$ or $h^{\text{DPL}}(x, \lambda)$ and it satisfies the following:

- (P.1) $h(x, \lambda)$ is a monotone increasing function of x ($x \in D$) and its range is \mathbb{R} , where D is the domain of transformation, namely $D = \mathbb{R}_+$ for DPT and $D = \mathbb{R}_{(0,1)}$ for DPLT.
- (P.2) $h(x, \lambda)$ and $h^{-1}(x, \lambda)$ are three times continuously differentiable, where $f(x, \lambda) = h^{-1}(x, \lambda)$ is the inverse function of $h(x, \lambda)$ defined by $x = h(f(x, \lambda), \lambda)$.

(P.3) The moments of the following exist for each fixed $\lambda > 0$:

- (1) $\{h_\lambda(x, \lambda)\}^2$ and $\{(\partial/\partial\lambda) \log(h_x(x, \lambda))\}^2$,
- (2) $\{h_\lambda(x, \lambda)\}^4$, $\{h_{\lambda\lambda}(x, \lambda)\}^2$ and $\{(\partial^2/\partial\lambda^2) \log(h_x(x, \lambda))\}^2$,
- (3) $h^{-1}(x, \lambda)$, $h_x^{-1}(x, \lambda)$, $h_\lambda^{-1}(x, \lambda)$, $h_{\lambda x}^{-1}(x, \lambda)$, $h_{xx}^{-1}(x, \lambda)$ and $h_{\lambda\lambda}^{-1}(x, \lambda)$,

where their expectations are taken with respect to $h(x, \lambda)$ which is normally distributed.

Property (P.1) means that the transformation is a one-to-one and onto function from D to \mathbb{R} . Clearly, (P.1) is not satisfied by the Box-Cox transformation (Box and Cox, 1964), but by the logarithmic transformation. Properties (P.2) and (P.3) will be used for establishing consistency of estimators including transformation parameter λ and for constructing prediction intervals. Especially, (P.2) and (P.3) (1) guarantees that the random variable $F(\hat{\beta}(\lambda), \hat{\sigma}_v^2(\lambda), \hat{\sigma}_e^2(\lambda), \lambda)$ given in (11) converges in probability, and (P.3)(2) guarantees that $(\partial/\partial\lambda)F(\hat{\beta}(\lambda), \hat{\sigma}_v^2(\lambda), \hat{\sigma}_e^2(\lambda), \lambda)$ converges in probability.

A.1 Proof of Lemma 1. Recall that $\hat{\beta}(\lambda) = \hat{\beta}(\hat{\rho}(\lambda), \lambda)$. It is noted that

$$\hat{\beta}(\lambda) - \beta = \hat{\beta}(\rho, \lambda) - \beta + \left(\frac{\partial \hat{\beta}(\rho, \lambda)}{\partial \rho} \right)' (\hat{\rho}(\lambda) - \rho) + O_p((\hat{\rho}(\lambda) - \rho)^2). \quad (30)$$

Here $\partial \hat{\beta}(\rho, \lambda)/\partial \rho$ is expressed as

$$\frac{\partial \hat{\beta}(\rho, \lambda)}{\partial \rho} = \left(\sum_{i=1}^m \frac{n_i \bar{\mathbf{x}}_i \bar{\mathbf{x}}_i'}{1 + n_i \rho} + \sum_{i=1}^m \sum_{j=1}^{n_i} \mathbf{x}_{ij} \mathbf{x}_{ij}' \right)^{-1} \left(\sum_{i=1}^m \frac{n_i^2 \bar{\mathbf{x}}_i \bar{\mathbf{x}}_i'}{(1 + n_i \rho)^2} \right) (\hat{\beta}(\rho, \lambda) - \hat{\beta}^\dagger(\rho, \lambda)),$$

where $\hat{\beta}^\dagger(\rho, \lambda) = (\sum_{i=1}^m n_i^2 \bar{\mathbf{x}}_i \bar{\mathbf{x}}_i' / (1 + n_i \rho)^2)^{-1} \sum_{i=1}^m n_i^2 \bar{\mathbf{x}}_i z_i(\lambda) / (1 + n_i \rho)^2$. Note that $(\hat{\beta}^\dagger(\rho, \lambda) - \hat{\beta}(\lambda)) | \mathbf{y}_m = O_p(m^{-1/2})$ since $(\hat{\beta}^\dagger(\rho, \lambda) - \beta) | \mathbf{y}_m = O_p(m^{-1/2})$ from (29) and Assumption 1. Thus, $(\partial \hat{\beta}(\rho, \lambda)/\partial \rho) | \mathbf{y}_m = O_p(m^{-1/2})$. Also, $\hat{\rho} - \rho$ can be expanded as

$$\hat{\rho}(\lambda) - \rho = \frac{1}{\sigma_e^2} (\hat{\sigma}_e^2(\lambda) - \sigma_e^2) - \frac{\sigma_v^2}{\sigma_e^4} (\hat{\sigma}_v^2(\lambda) - \sigma_v^2) + O_p(m^{-1}), \quad (31)$$

which implies that $(\hat{\rho}(\lambda) - \rho) | \mathbf{y}_m = O_p(m^{-1/2})$ and $E[\hat{\rho}(\lambda) - \rho | \mathbf{y}_m] = O_p(m^{-1})$ from Assumptions (A.3) and (A.4). Combining these observations and applying (29) to the first term in the r.h.s. of (30), one gets $(\hat{\beta}(\lambda) - \beta) | \mathbf{y}_m = O_p(m^{-1/2})$ under Assumptions 1 and 2.

To show $E[\hat{\beta}(\lambda) - \beta | \mathbf{y}_m] = O_p(m^{-1/2})$, from (30), it is sufficient to show that $E[\hat{\beta}(\rho, \lambda) - \beta | \mathbf{y}_m] = O_p(m^{-1})$. Note that $\hat{\beta}(\rho, \lambda) - \beta$ is rewritten as

$$\begin{aligned} & \hat{\beta}(\rho, \lambda) - \beta \\ &= \left(\sum_{i=1}^m \frac{n_i \bar{\mathbf{x}}_i \bar{\mathbf{x}}_i'}{1 + n_i \rho} + \sum_{i=1}^m \sum_{j=1}^{n_i} \mathbf{x}_{ij} \mathbf{x}_{ij}' \right)^{-1} \left(\sum_{i=1}^m \frac{n_i \bar{\mathbf{x}}_i (z_i(\lambda) - \bar{\mathbf{x}}_i' \beta)}{1 + n_i \rho} + \sum_{i=1}^m \sum_{j=1}^{n_i} \mathbf{x}_{ij} (h(y_{ij}, \lambda) - \mathbf{x}_{ij}' \beta) \right) \\ &= \left(\sum_{i=1}^m \frac{n_i \bar{\mathbf{x}}_i \bar{\mathbf{x}}_i'}{1 + n_i \rho} + \sum_{i=1}^m \sum_{j=1}^{n_i} \mathbf{x}_{ij} \mathbf{x}_{ij}' \right)^{-1} \left(\sum_{i=1}^{m-1} \frac{n_i \bar{\mathbf{x}}_i (z_i(\lambda) - \bar{\mathbf{x}}_i' \beta)}{1 + n_i \rho} + \sum_{i=1}^{m-1} \sum_{j=1}^{n_i} \mathbf{x}_{ij} (h(y_{ij}, \lambda) - \mathbf{x}_{ij}' \beta) \right) \\ &\quad + \left(\sum_{i=1}^m \frac{n_i \bar{\mathbf{x}}_i \bar{\mathbf{x}}_i'}{1 + n_i \rho} + \sum_{i=1}^m \sum_{j=1}^{n_i} \mathbf{x}_{ij} \mathbf{x}_{ij}' \right)^{-1} \left(\frac{n_m \bar{\mathbf{x}}_m (z_m(\lambda) - \bar{\mathbf{x}}_m' \beta)}{1 + n_m \rho} + \sum_{j=1}^{n_m} \mathbf{x}_{mj} (h(y_{mj}, \lambda) - \mathbf{x}_{mj}' \beta) \right). \end{aligned}$$

Noting that $z_1(\lambda), \dots, z_{m-1}(\lambda)$ are independent of \mathbf{y}_m , $z_i(\lambda) = O_p(1)$, $\sum_{i=1}^m n_i \bar{\mathbf{x}}_i \bar{\mathbf{x}}'_i / (1 + n_i \rho) = \mathbf{O}(m)$ and $\sum_{i=1}^m \sum_{j=1}^{n_i} \mathbf{x}_{ij} \mathbf{x}'_{ij} = \mathbf{O}(m)$ under Assumption 1, one gets

$$\begin{aligned} E[\hat{\boldsymbol{\beta}}(\rho, \lambda) - \boldsymbol{\beta} | \mathbf{y}_m] &= \left(\sum_{i=1}^m \frac{n_i \bar{\mathbf{x}}_i \bar{\mathbf{x}}'_i}{1 + n_i \rho} + \sum_{i=1}^m \sum_{j=1}^{n_i} \mathbf{x}_{ij} \mathbf{x}'_{ij} \right)^{-1} \left(\frac{n_m \bar{\mathbf{x}}_m (z_m(\lambda) - \bar{\mathbf{x}}'_m \boldsymbol{\beta})}{1 + n_m \rho} + \sum_{j=1}^{n_i} \mathbf{x}_{mj} (h(y_{mj}, \lambda) - \mathbf{x}'_{mj} \boldsymbol{\beta}) \right) \\ &= \mathbf{O}_p(m^{-1}). \end{aligned}$$

To show the third part, by straightforward calculation, one gets

$$\begin{aligned} \frac{\partial \hat{\boldsymbol{\beta}}(\lambda)}{\partial \lambda} &= \left(\sum_{i=1}^m \frac{n_i \bar{\mathbf{x}}_i \bar{\mathbf{x}}'_i}{1 + n_i \hat{\rho}(\lambda)} + \sum_{i=1}^m \sum_{j=1}^{n_i} \mathbf{x}_{ij} \mathbf{x}'_{ij} \right)^{-1} \sum_{i=1}^m \frac{n_i \bar{\mathbf{x}}_i \bar{\mathbf{x}}'_i}{(1 + n_i \hat{\rho}(\lambda))^2} \{ \hat{\boldsymbol{\beta}}(\lambda) - \hat{\boldsymbol{\beta}}^\dagger(\lambda) \} \left(\frac{\partial \hat{\rho}(\lambda)}{\partial \lambda} \right) \\ &\quad + \left(\sum_{i=1}^m \frac{n_i \bar{\mathbf{x}}_i \bar{\mathbf{x}}'_i}{1 + n_i \hat{\rho}(\lambda)} + \sum_{i=1}^m \sum_{j=1}^{n_i} \mathbf{x}_{ij} \mathbf{x}'_{ij} \right)^{-1} \left(\sum_{i=1}^m \frac{n_i \bar{\mathbf{x}}_i z_{i,\lambda}(\lambda)}{1 + n_i \hat{\rho}(\lambda)} + \sum_{i=1}^m \sum_{j=1}^{n_i} \mathbf{x}_{ij} h_\lambda(y_{ij}, \lambda) \right), \end{aligned}$$

where $\hat{\boldsymbol{\beta}}^\dagger(\lambda) = (\sum_{i=1}^m n_i^2 \bar{\mathbf{x}}_i \bar{\mathbf{x}}'_i / (1 + n_i \hat{\rho}(\lambda))^2)^{-1} \sum_{i=1}^m n_i^2 \bar{\mathbf{x}}_i z_i(\lambda) / (1 + n_i \hat{\rho}(\lambda))^2$,

$$\frac{\partial \hat{\rho}(\lambda)}{\partial \lambda} = \frac{1}{\hat{\sigma}_e^2(\lambda)} \left(\frac{\partial}{\partial \lambda} \hat{\sigma}_v^2(\lambda) - \hat{\rho}(\lambda) \frac{\partial}{\partial \lambda} \hat{\sigma}_e^2(\lambda) \right) \quad \text{and} \quad z_{i,\lambda} = \frac{\partial}{\partial \lambda} z_i(\lambda), \quad i = 1, \dots, m.$$

Since $\hat{\boldsymbol{\beta}}^\dagger(\lambda) - \boldsymbol{\beta} = \mathbf{O}_p(m^{-1/2})$, we have $\hat{\boldsymbol{\beta}}(\lambda) - \hat{\boldsymbol{\beta}}^\dagger(\lambda) = \mathbf{O}_p(m^{-1/2})$. Also note that $\partial \hat{\rho}(\lambda) / \partial \lambda | \mathbf{y}_m = O_p(1)$ from Assumption (A.8). Then, we have

$$\begin{aligned} E\left[\frac{\partial \hat{\boldsymbol{\beta}}(\lambda)}{\partial \lambda} \middle| \mathbf{y}_m \right] &= \left(\sum_{i=1}^m \frac{n_i \bar{\mathbf{x}}_i \bar{\mathbf{x}}'_i}{1 + n_i \rho} + \sum_{i=1}^m \sum_{j=1}^{n_i} \mathbf{x}_{ij} \mathbf{x}'_{ij} \right)^{-1} \left(\sum_{i=1}^{m-1} \frac{n_i \bar{\mathbf{x}}_i E[z_i(\lambda)]}{1 + n_i \rho} + \sum_{i=1}^{m-1} \sum_{j=1}^{n_i} \mathbf{x}_{ij} E[h_\lambda(y_{ij}, \lambda)] \right) \\ &\quad + \left(\sum_{i=1}^m \frac{n_i \bar{\mathbf{x}}_i \bar{\mathbf{x}}'_i}{1 + n_i \rho} + \sum_{i=1}^m \sum_{j=1}^{n_i} \mathbf{x}_{ij} \mathbf{x}'_{ij} \right)^{-1} \left(\frac{n_m \bar{\mathbf{x}}_m z_m(\lambda)}{1 + n_m \rho} + \sum_{j=1}^{n_m} \mathbf{x}_{mj} h_\lambda(y_{mj}, \lambda) \right) + \mathbf{O}_p(m^{-1/2}). \end{aligned}$$

Hence, one gets

$$\begin{aligned} \sqrt{m} \left(\frac{\partial \hat{\boldsymbol{\beta}}(\lambda)}{\partial \lambda} - E\left[\frac{\partial \hat{\boldsymbol{\beta}}(\lambda)}{\partial \lambda} \middle| \mathbf{y}_m \right] \right) \bigg| \mathbf{y}_m &= \left(\frac{1}{m} \sum_{i=1}^m \frac{n_i \bar{\mathbf{x}}_i \bar{\mathbf{x}}'_i}{1 + n_i \rho} + \frac{1}{m} \sum_{i=1}^m \sum_{j=1}^{n_i} \mathbf{x}_{ij} \mathbf{x}'_{ij} \right)^{-1} \frac{\sqrt{m-1}}{\sqrt{m}} \frac{1}{\sqrt{m-1}} \\ &\quad \times \left(\sum_{i=1}^{m-1} \frac{n_i \bar{\mathbf{x}}_i}{1 + n_i \rho} \{ z_i(\lambda) - E[z_i(\lambda)] \} + \sum_{i=1}^{m-1} \sum_{j=1}^{n_i} \mathbf{x}_{ij} \{ h_\lambda(y_{ij}, \lambda) - E[h_\lambda(y_{ij}, \lambda)] \} \right) + \mathbf{O}_p(1), \end{aligned}$$

which is of order $\mathbf{O}_p(1)$, since it follows from CLT and Assumption 1,

$$\frac{1}{\sqrt{m-1}} \sum_{i=1}^{m-1} \frac{n_i \bar{\mathbf{x}}_i}{1 + n_i \rho} \{ z_i(\lambda) - E[z_i(\lambda)] \} = \mathbf{O}_p(1).$$

and

$$\frac{1}{\sqrt{m-1}} \sum_{i=1}^{m-1} \sum_{j=1}^{n_i} \mathbf{x}_{ij} \{h_\lambda(y_{ij}, \lambda) - E[h_\lambda(y_{ij}, \lambda)]\} = \mathbf{O}_p(1)$$

Therefore, Lemma 1 is proved. \square

A.2 Proof of Lemma 2. We can easily verify that Assumptions (A.3) and (A.4) are satisfied for the three estimators of $\boldsymbol{\sigma}^2$ based on their stochastic expansions given in Prasad and Rao (1990), Datta and Lahiri (2000) and Das, Jiang and Rao (2004). Thus, we shall check Assumptions (A.8) and (A.6) for $i = m$.

[1] **PR estimator.** Recall that $\widehat{\boldsymbol{\sigma}}_{PR}^2$ is given in (8). For S_1 ,

$$\begin{aligned} \frac{1}{m} \left(\frac{\partial}{\partial \lambda} S_1(\lambda) \right) &= \frac{2}{m} \mathbf{h}(\mathbf{Y}, \lambda)' (\mathbf{I}_N - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}') \left(\frac{\partial}{\partial \lambda} \mathbf{h}(\mathbf{Y}, \lambda) \right) \\ &= \frac{2}{m} \sum_{i=1}^m \mathbf{h}(\mathbf{y}_i, \lambda)' \mathbf{M}_i \left(\frac{\partial}{\partial \lambda} \mathbf{h}(\mathbf{y}_i, \lambda) \right) + o_p(1) \\ &= \frac{2}{m} \sum_{i=1}^{m-1} \mathbf{h}(\mathbf{y}_i, \lambda)' \mathbf{M}_i \left(\frac{\partial}{\partial \lambda} \mathbf{h}(\mathbf{y}_i, \lambda) \right) + \frac{2}{m} \mathbf{h}(\mathbf{y}_m, \lambda)' \mathbf{M}_m \left(\frac{\partial}{\partial \lambda} \mathbf{h}(\mathbf{y}_m, \lambda) \right) + o_p(1), \end{aligned}$$

where $\mathbf{M}_i = \mathbf{I}_{n_i} - \mathbf{X}_i(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'_i$. Then from (28) and (29), it follows that

$$\frac{1}{m} E \left[\frac{\partial}{\partial \lambda} S_1(\lambda) \middle| \mathbf{y}_m \right] = O_p(1) \quad \text{and} \quad \frac{1}{m} \left(\frac{\partial}{\partial \lambda} S_1(\lambda) - \frac{1}{m} E \left[\frac{\partial}{\partial \lambda} S_1(\lambda) \middle| \mathbf{y}_m \right] \right) \middle| \mathbf{y}_m = O_p(m^{-1/2}).$$

For S_2 , we can show similar properties since

$$\frac{\partial}{\partial \lambda} S_2(\lambda) = 2 \mathbf{h}(\mathbf{Y}, \lambda)' (\mathbf{E} - \mathbf{E}\mathbf{X}(\mathbf{X}'\mathbf{E}\mathbf{X})^{-1}\mathbf{X}'\mathbf{E}) \left(\frac{\partial}{\partial \lambda} \mathbf{h}(\mathbf{Y}, \lambda) \right).$$

Thus, Assumptions (A.8) and (A.6) are satisfied.

[2] **ML.** The ML estimator $\widehat{\boldsymbol{\sigma}}_{ML}^2$ is given in (9). From the implicit function theorem,

$$\frac{\partial}{\partial \lambda} \widehat{\boldsymbol{\sigma}}_{ML}^2(\lambda) = \mathbf{I}(\lambda)^{-1} \mathbf{J}(\lambda),$$

where

$$\begin{aligned} \mathbf{I}(\lambda) &= \begin{pmatrix} I_{11}(\lambda) & I_{12}(\lambda) \\ I_{21}(\lambda) & I_{22}(\lambda) \end{pmatrix} = \begin{pmatrix} \partial L_1(\boldsymbol{\sigma}^2, \lambda) / \partial \sigma_e^2 & \partial L_1(\boldsymbol{\sigma}^2, \lambda) / \partial \sigma_v^2 \\ \partial L_2(\boldsymbol{\sigma}^2, \lambda) / \partial \sigma_e^2 & \partial L_2(\boldsymbol{\sigma}^2, \lambda) / \partial \sigma_v^2 \end{pmatrix} \Big|_{\boldsymbol{\sigma}^2 = \widehat{\boldsymbol{\sigma}}_{ML}^2(\lambda)}, \\ \mathbf{J}(\lambda) &= (J_1(\lambda), J_2(\lambda))' = \left(\frac{\partial}{\partial \lambda} L_1(\boldsymbol{\sigma}^2, \lambda) \Big|_{\boldsymbol{\sigma}^2 = \widehat{\boldsymbol{\sigma}}_{ML}^2(\lambda)}, \frac{\partial}{\partial \lambda} L_2(\boldsymbol{\sigma}^2, \lambda) \Big|_{\boldsymbol{\sigma}^2 = \widehat{\boldsymbol{\sigma}}_{ML}^2(\lambda)} \right)'. \end{aligned} \quad (32)$$

By straightforward calculation, it is shown that

$$\begin{aligned} \frac{\partial}{\partial \lambda} L_1(\boldsymbol{\sigma}^2, \lambda) &= \frac{1}{\sigma_e^4} \sum_{i=1}^m \left(\mathbf{h}(\mathbf{y}_i, \lambda) - \mathbf{X}_i \widehat{\boldsymbol{\beta}}(\rho, \lambda) - \frac{n_i \rho}{1 + n_i \rho} (z_i(\lambda) - \bar{\mathbf{x}}_i' \widehat{\boldsymbol{\beta}}(\rho, \lambda)) \mathbf{j}_i \right)' \\ &\quad \cdot \left(\mathbf{X}_i + \frac{n_i \rho}{1 + n_i \rho} \mathbf{j}_i \bar{\mathbf{x}}_i' \right) \left(\frac{\partial}{\partial \lambda} \widehat{\boldsymbol{\beta}}(\rho, \lambda) \right), \end{aligned}$$

where

$$\frac{\partial}{\partial \lambda} \hat{\beta}(\rho, \lambda) = \left(\sum_{i=1}^m \frac{n_i \bar{\mathbf{x}}_i \bar{\mathbf{x}}_i'}{1 + n_i \rho} \right)^{-1} \sum_{i=1}^m \frac{n_i \bar{\mathbf{x}}_i z_{i,\lambda}(\lambda)}{1 + n_i \rho},$$

which is of order $\mathbf{O}_p(1)$ under Assumption 1 and 2. Then from the above expression, it easily follows that

$$\begin{aligned} \frac{1}{m} J_1(\lambda) &= \sigma_e^{-4} \frac{1}{m} \sum_{i=1}^m \left(\mathbf{h}(\mathbf{y}_i, \lambda) - \mathbf{X}_i \boldsymbol{\beta} - \frac{n_i \rho}{1 + n_i \rho} (z_i(\lambda) - \bar{\mathbf{x}}_i' \boldsymbol{\beta}) \mathbf{j}_i \right)' \\ &\quad \cdot \left(\mathbf{X}_i + \frac{n_i \rho}{1 + n_i \rho} \mathbf{j}_i \bar{\mathbf{x}}_i' \right) \left(\frac{\partial}{\partial \lambda} \hat{\beta}(\rho, \lambda) \right) + O_p(m^{-1/2}) \\ &= \left(\frac{1}{m} \sum_{i=1}^m \mathbf{J}_{1i}(\lambda) \right)' \left(\frac{\partial}{\partial \lambda} \hat{\beta}(\rho, \lambda) \right) + O_p(m^{-1/2}), \quad (\text{say}). \end{aligned}$$

Since $\mathbf{J}_{1i}(\lambda)$, $i = 1, \dots, m$, are mutually independent random vectors $E[\mathbf{J}_{1i}(\lambda)] = \mathbf{0}$, it is seen that $m^{-1} \sum_{i=1}^m \mathbf{J}_{1i}(\lambda) | \mathbf{y}_m = O_p(m^{-1/2})$ from (29). Thus, $m^{-1} J_1(\lambda) | \mathbf{y}_m = O_p(m^{-1/2})$, namely,

$$m^{-1/2} J_1(\lambda) | \mathbf{y}_m = O_p(1).$$

Also, we obtain

$$\frac{\partial}{\partial \lambda} L_2(\boldsymbol{\sigma}^2, \lambda) = \sum_{i=1}^m \frac{2n_i^2}{(\sigma_e^2 + n_i \sigma_v^2)^2} \left(z_i(\lambda) - \bar{\mathbf{x}}_i' \hat{\beta}(\rho, \lambda) \right) \left(z_{i,\lambda}(\lambda) - \bar{\mathbf{x}}_i' \left(\frac{\partial}{\partial \lambda} \hat{\beta}(\rho, \lambda) \right) \right),$$

which implies that

$$\begin{aligned} \frac{1}{m} J_2(\lambda) &= \frac{1}{m} \sum_{i=1}^m \frac{2n_i^2}{(\sigma_e^2 + n_i \sigma_v^2)^2} (z_i(\lambda) - \bar{\mathbf{x}}_i' \boldsymbol{\beta}) z_{i,\lambda}(\lambda) + O_p(m^{-1/2}) \\ &= \frac{1}{m} \sum_{i=1}^m J_{2i}(\lambda) + O_p(m^{-1/2}), \quad (\text{say}). \end{aligned}$$

Since $J_{2i}(\lambda) = O_p(1)$ and it depends only on \mathbf{y}_i of \mathbf{Y} , from (28) and (29), one gets

$$m^{-1} J_2(\lambda) | \mathbf{y}_m = O_p(1) \quad \text{and} \quad m^{-1/2} (J_2(\lambda) - E[J_2(\lambda) | \mathbf{y}_m]) | \mathbf{y}_m = O_p(m^{-1/2}).$$

We next evaluate $\mathbf{I}(\lambda)$. We here give a proof for $I_{21}(\lambda)$, and we omit proofs for the other elements since they can be similarly proved. By a straightforward calculation,

$$\begin{aligned} \frac{\partial L_2(\boldsymbol{\sigma}^2, \lambda)}{\partial \sigma_v^2} &= - \sum_{i=1}^m \frac{2n_i^2}{(\sigma_e^2 + n_i \sigma_v^2)^3} (z_i(\lambda) - \bar{\mathbf{x}}_i' \hat{\beta}(\rho, \lambda))^2 + \sum_{i=1}^m \frac{2n_i^2}{(\sigma_e^2 + n_i \sigma_v^2)^3} \\ &\quad - \sum_{i=1}^m \frac{2n_i^2}{(\sigma_e^2 + n_i \sigma_v^2)^2} (z_i(\lambda) - \bar{\mathbf{x}}_i' \hat{\beta}(\rho, \lambda)) \bar{\mathbf{x}}_i' \left(\frac{\partial \hat{\beta}(\rho, \lambda)}{\partial \sigma_e^2} \right), \end{aligned}$$

where

$$\frac{\partial \hat{\beta}(\rho, \lambda)}{\partial \sigma_e^2} = - \frac{\sigma_v^2}{\sigma_e^4} \left(\sum_{i=1}^m \frac{n_i \bar{\mathbf{x}}_i \bar{\mathbf{x}}_i'}{1 + n_i \rho} \right)^{-1} \sum_{i=1}^m \frac{n_i \bar{\mathbf{x}}_i \bar{\mathbf{x}}_i'}{(1 + n_i \rho)^2} (\hat{\beta}(\rho, \lambda) - \hat{\beta}^\dagger(\rho, \lambda)),$$

which is $\mathbf{O}_p(m^{-1/2})$ since $\widehat{\boldsymbol{\beta}}(\rho, \lambda) - \widehat{\boldsymbol{\beta}}^\dagger(\rho, \lambda) = \mathbf{O}_p(m^{-1/2})$ as in the proof of Lemma 1. Then,

$$\frac{1}{m}I_{21}(\lambda) = -\frac{1}{m} \sum_{i=1}^m \frac{2n_i^2}{(\sigma_e^2 + n_i\sigma_v^2)^3} (z_i(\lambda) - \bar{\mathbf{x}}_i\boldsymbol{\beta})^2 + \frac{1}{m} \sum_{i=1}^m \frac{2n_i^2}{(\sigma_e^2 + n_i\sigma_v^2)^3} + O_p(m^{-1/2}),$$

so that, we have

$$\begin{aligned} \frac{1}{m}I_{21}(\boldsymbol{\sigma}^2, \lambda) \Big| \mathbf{y}_m &= -\frac{1}{m} \sum_{i=1}^{m-1} \frac{2n_i^2}{(\sigma_e^2 + n_i\sigma_v^2)^2} + \frac{1}{m} \sum_{i=1}^m \frac{2n_i^2}{(\sigma_e^2 + n_i\sigma_v^2)^3} \\ &\quad - \frac{1}{m} \frac{2n_i^2}{(\sigma_e^2 + n_m\sigma_v^2)^3} (z_m(\lambda) - \bar{\mathbf{x}}_m\boldsymbol{\beta})^2 + O_p(m^{-1/2}) \\ &= -\frac{1}{m} \sum_{i=1}^{m-1} \frac{2n_i^2}{(\sigma_e^2 + n_i\sigma_v^2)^2} + \frac{1}{m} \sum_{i=1}^m \frac{2n_i^2}{(\sigma_e^2 + n_i\sigma_v^2)^3} + O_p(m^{-1/2}), \end{aligned}$$

since $E[(z_i(\lambda) - \bar{\mathbf{x}}_i\boldsymbol{\beta})^2] = \sigma_e^2 + n_i\sigma_v^2$. This demonstrates that the leading term is of order $O(1)$. Since the other elements of $\mathbf{I}(\lambda)$ can be evaluated similarly, we have

$$m^{-1}\mathbf{I}(\lambda)|\mathbf{y}_i = \mathbf{C}(\boldsymbol{\theta}) + \mathbf{O}_p(m^{-1/2}),$$

where $\mathbf{C}(\boldsymbol{\theta})$ is a non-stochastic matrix with bounded entries, i.e. $\mathbf{C}(\boldsymbol{\theta}) = \mathbf{O}(1)$. Therefore, one gets

$$\frac{\partial}{\partial \lambda} \widehat{\boldsymbol{\sigma}}_{ML}^2(\lambda) \Big| \mathbf{y}_m = (m^{-1}\mathbf{I}(\lambda))^{-1} m^{-1} \mathbf{J}(\lambda) \Big| \mathbf{y}_m = \mathbf{O}_p(1),$$

which shows that the ML estimator satisfies Assumption (A.8). Moreover,

$$\begin{aligned} \sqrt{m} \left(\frac{\partial}{\partial \lambda} \widehat{\boldsymbol{\sigma}}_{ML}^2(\lambda) - E \left[\frac{\partial}{\partial \lambda} \widehat{\boldsymbol{\sigma}}_{ML}^2(\lambda) \Big| \mathbf{y}_m \right] \right) \Big| \mathbf{y}_m \\ = \mathbf{C}(\boldsymbol{\theta})^{-1} m^{-1/2} (\mathbf{J}(\lambda) - E[\mathbf{J}(\lambda) | \mathbf{y}_m]) \Big| \mathbf{y}_m + \mathbf{O}_p(1), \end{aligned}$$

which is of order $\mathbf{O}_p(1)$. Thus, (A.6) is satisfied.

[3] REML Recall that REML is given in (10). From the implicit function theorem,

$$\frac{\partial}{\partial \lambda} \widehat{\boldsymbol{\sigma}}_{RML}^2(\lambda) = \mathbf{I}^R(\lambda)^{-1} \mathbf{J}(\lambda),$$

where $\mathbf{J}(\lambda)$ is defined in (32) and

$$\mathbf{I}^R(\lambda) = \begin{pmatrix} \mathbf{I}_{11}^R(\lambda) & \mathbf{I}_{12}^R(\lambda) \\ \mathbf{I}_{21}^R(\lambda) & \mathbf{I}_{22}^R(\lambda) \end{pmatrix} = \mathbf{I}(\lambda) + \begin{pmatrix} \partial P_1(\boldsymbol{\sigma}^2)/\partial \sigma_e^2 & \partial P_1(\boldsymbol{\sigma}^2)/\partial \sigma_v^2 \\ \partial P_2(\boldsymbol{\sigma}^2)/\partial \sigma_e^2 & \partial P_2(\boldsymbol{\sigma}^2)/\partial \sigma_v^2 \end{pmatrix} \Big|_{\boldsymbol{\sigma}^2 = \widehat{\boldsymbol{\sigma}}_{RML}^2(\lambda)}.$$

for $P_1(\boldsymbol{\sigma}^2) = \text{tr}[(\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Sigma}^{-2}\mathbf{X}]$ and $P_2(\boldsymbol{\sigma}^2) = \text{tr}[(\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{Z}\boldsymbol{\Sigma}^{-1}\mathbf{X}]$. Then, the result follows if $m^{-1}P_1(\boldsymbol{\sigma}^2)|\mathbf{y}_i = O_p(1)$ and $m^{-1}P_2(\boldsymbol{\sigma}^2)|\mathbf{y}_i = O_p(1)$, which can be seen from Assumptions 1 and (A.3). \square

A.3 Proof of Lemma 3. We begin by demonstrating the consistency of $\hat{\lambda}$. According the Cramer method explained in Jiang (2010), we show that the equation $F(\hat{\beta}(\lambda), \hat{\sigma}_v^2(\lambda), \hat{\sigma}_e^2(\lambda), \lambda) = 0$ includes a solution which converges to λ in probability. Let

$$g_m(\lambda') = m^{-1}F(\hat{\beta}(\lambda'), \hat{\sigma}_v^2(\lambda'), \hat{\sigma}_e^2(\lambda'), \lambda'),$$

for scalar λ' . Then, it can be seen that $g_m(\lambda')$ converges to $g(\lambda')$ in probability, where

$$g(\lambda') = \lim_{m \rightarrow \infty} m^{-1}E_\lambda[F(\hat{\beta}(\lambda'), \hat{\sigma}_v^2(\lambda'), \hat{\sigma}_e^2(\lambda'), \lambda')].$$

When $\lambda' = \lambda$, it is noted that $g(\lambda) = 0$, since $g(\lambda) = \lim_{m \rightarrow \infty} m^{-1}E_\lambda[F(\beta, \sigma_v^2, \sigma_e^2, \lambda)] = 0$. Since $g(\lambda')$ is continuous, without loss of generality, we have $g(\lambda - \varepsilon) < 0$ and $g(\lambda + \varepsilon) > 0$ for some positive ε . Then, $g_m(\lambda - \varepsilon)$ and $g_m(\lambda + \varepsilon)$ converge to $g(\lambda - \varepsilon) < 0$ and $g(\lambda + \varepsilon) > 0$, respectively, in probability. This implies that both probabilities $P(g_m(\lambda - \varepsilon) < 0)$ and $P(g_m(\lambda + \varepsilon) > 0)$ converge to one as $m \rightarrow \infty$. In fact, for instance, the former result follows from the fact that

$$\begin{aligned} P(g_m(\lambda - \varepsilon) < 0) &= P(g_m(\lambda - \varepsilon) - g(\lambda - \varepsilon) < -g(\lambda - \varepsilon)) \\ &> P(|g_m(\lambda - \varepsilon) - g(\lambda - \varepsilon)| < -g(\lambda - \varepsilon)) \rightarrow 1, \end{aligned}$$

as $m \rightarrow \infty$ since $-g(\lambda - \varepsilon) > 0$. Thus, for any $\delta > 0$, there exists an M such that for any $m > M$, $P(g_m(\lambda - \varepsilon) < 0) > 1 - \delta$ and $P(g_m(\lambda + \varepsilon) > 0) > 1 - \delta$. Note that the intersection of the events $\{g_m(\lambda - \varepsilon) < 0\}$ and $\{g_m(\lambda + \varepsilon) > 0\}$ implies that $\hat{\lambda}$ is included in the interval $(\lambda - \varepsilon, \lambda + \varepsilon)$, namely, $|\hat{\lambda} - \lambda| < \varepsilon$. Hence,

$$P(|\hat{\lambda} - \lambda| < \varepsilon) > P(g_m(\lambda - \varepsilon) < 0, g_m(\lambda + \varepsilon) > 0) > 1 - 2\delta,$$

which means that $\hat{\lambda}$ is consistent.

We next show that $(\hat{\lambda} - \lambda)|\mathbf{y}_m = O_p(m^{-1/2})$ in the case of $i = m$. To this end, we expand the equation (11) around λ to get

$$\sqrt{m}(\hat{\lambda} - \lambda) = -\frac{m^{-1/2}F(\hat{\beta}(\lambda), \hat{\sigma}_v^2(\lambda), \hat{\sigma}_e^2(\lambda), \lambda)}{m^{-1}\left(\frac{\partial}{\partial \lambda}F(\hat{\beta}(\lambda), \hat{\sigma}_v^2(\lambda), \hat{\sigma}_e^2(\lambda), \lambda)\right)|_{\lambda=\lambda^*}}, \quad (33)$$

where λ^* is an intermediate value between λ and $\hat{\lambda}$. For the numerator in (33), from Lemma 1 and Assumption 2, it is seen that

$$\begin{aligned} \frac{1}{\sqrt{m}}F(\hat{\beta}(\lambda), \hat{\sigma}_v^2(\lambda), \hat{\sigma}_e^2(\lambda), \lambda)|\mathbf{y}_m &= \frac{1}{\sqrt{m}}F(\beta, \sigma_v^2, \sigma_e^2, \lambda)|\mathbf{y}_m + O_p(1) \\ &= \frac{1}{\sqrt{m}}\sum_{i=1}^m F_i(\theta)|\mathbf{y}_m + O_p(1), \end{aligned}$$

where $F_i(\theta) = -2\log f(\mathbf{y}_i; \theta)$ for $\theta = (\beta, \sigma_v^2, \sigma_e^2, \lambda)$ and the density function $f(\mathbf{y}_i; \theta)$ of \mathbf{y}_i . Since F_1, \dots, F_m are mutually independently distributed with $E[F_i(\theta)] = 0$, from (29), it is seen that

$$\frac{1}{\sqrt{m}}F(\hat{\beta}(\lambda), \hat{\sigma}_v^2(\lambda), \hat{\sigma}_e^2(\lambda), \lambda)|\mathbf{y}_m = O_p(1).$$

For the denominator in (33), it follows from the consistency of $\widehat{\lambda}$ that

$$m^{-1} \left(\frac{\partial}{\partial \lambda} F(\widehat{\beta}(\lambda), \widehat{\sigma}_v^2(\lambda), \widehat{\sigma}_e^2(\lambda), \lambda) \Big|_{\lambda=\lambda^*} \right) = m^{-1} \frac{\partial}{\partial \lambda} F(\widehat{\beta}(\lambda), \widehat{\sigma}_v^2(\lambda), \widehat{\sigma}_e^2(\lambda), \lambda) (1 + o_p(1)).$$

By straightforward calculation, it can be seen from Lemma 1 and Assumption 2 that

$$\begin{aligned} & \frac{1}{m} \left(\frac{\partial}{\partial \lambda} F(\widehat{\beta}(\lambda), \widehat{\sigma}_v^2(\lambda), \widehat{\sigma}_e^2(\lambda), \lambda) \right) \\ &= \left\{ -\frac{1}{\sigma_e^4} \left(\frac{\partial}{\partial \lambda} \widehat{\sigma}_e^2(\lambda) \right) \frac{1}{m} \sum_{i=1}^m (\mathbf{h}(\mathbf{y}_i, \lambda) - \mathbf{X}_i \boldsymbol{\beta})' \mathbf{V}_i(\rho)^{-1} \left(\frac{\partial}{\partial \lambda} \mathbf{h}(\mathbf{y}_i, \lambda) \right) \right. \\ &+ \sigma_e^{-2} \frac{1}{m} \sum_{i=1}^m \left(\frac{\partial}{\partial \lambda} \mathbf{h}(\mathbf{y}_i, \lambda) - \mathbf{X}_i \left(\frac{\partial}{\partial \lambda} \widehat{\beta}(\lambda) \right) \right)' \mathbf{V}_i(\rho)^{-1} \left(\frac{\partial}{\partial \lambda} \mathbf{h}(\mathbf{y}_i, \lambda) \right) \\ &+ \sigma_e^{-2} \frac{1}{m} \sum_{i=1}^m (\mathbf{h}(\mathbf{y}_i, \lambda) - \mathbf{X}_i \boldsymbol{\beta})' \left(\frac{\partial}{\partial \lambda} \mathbf{V}_i(\widehat{\rho}(\lambda))^{-1} \right) \left(\frac{\partial}{\partial \lambda} \mathbf{h}(\mathbf{y}_i, \lambda) \right) \\ &+ \left. \sigma_e^{-2} \frac{1}{m} \sum_{i=1}^m (\mathbf{h}(\mathbf{y}_i, \lambda) - \mathbf{X}_i \boldsymbol{\beta})' \mathbf{V}_i(\rho)^{-1} \left(\frac{\partial^2}{\partial \lambda^2} \mathbf{h}(\mathbf{y}_i, \lambda) \right) + \frac{1}{m} \left(\frac{\partial}{\partial \lambda} J(\mathbf{Y}, \lambda) \right) \right\} (1 + o_p(1)) \\ &= (K_1 + K_2 + K_3 + K_4 + K_5)(1 + o_p(1)). \quad (\text{say}) \end{aligned}$$

We shall evaluate the terms K_1, \dots, K_5 under Assumption 1. It is easy to see that $K_4|\mathbf{y}_m = O_p(1)$ and $K_5|\mathbf{y}_m = O_p(1)$ by (28). Similarly under Assumptions 1 and 2, we have $K_1 = O_p(1)$ by (28). To evaluate K_2 , the expression is rewritten as

$$\begin{aligned} K_2 &= \sigma_e^{-2} \frac{1}{m} \sum_{i=1}^m \left(\frac{\partial}{\partial \lambda} \mathbf{h}(\mathbf{y}_i, \lambda) \right)' \mathbf{V}_i(\rho)^{-1} \left(\frac{\partial}{\partial \lambda} \mathbf{h}(\mathbf{y}_i, \lambda) \right) \\ &+ \sigma_e^{-2} \left(\frac{\partial}{\partial \lambda} \widehat{\beta}(\lambda) \right)' \left(\frac{1}{m} \sum_{i=1}^m \mathbf{X}_i \mathbf{V}_i(\rho)^{-1} \left(\frac{\partial}{\partial \lambda} \mathbf{h}(\mathbf{y}_i, \lambda) \right) \right), \end{aligned}$$

where $(\partial \widehat{\beta}(\lambda)/\partial \lambda)|\mathbf{y}_m = O_p(1)$ from Lemma 1. Then from (28), $K_2|\mathbf{y}_m = O_p(1)$. For K_3 , it is observed that the each element of $(\partial \mathbf{V}_i(\widehat{\rho}(\lambda))^{-1}/\partial \lambda)|\mathbf{y}_m$ is of order $O_p(1)$, since $(\widehat{\rho}(\lambda)/\partial \lambda)|\mathbf{y}_m = O_p(1)$ under Assumption 2. Furthermore, the expression of K_3 reduces to $K_3 = \sigma_e^{-2} \text{tr}[\mathbf{K}_3^*]$, where

$$\mathbf{K}_3^* = \frac{1}{m} \sum_{i=1}^m \left(\frac{\partial}{\partial \lambda} \mathbf{h}(\mathbf{y}_i, \lambda) \right) (\mathbf{h}(\mathbf{y}_i, \lambda) - \mathbf{X}_i \boldsymbol{\beta})' \cdot \left(\frac{\partial}{\partial \lambda} \mathbf{V}_i(\widehat{\rho}(\lambda))^{-1} \right).$$

From (28) and Assumption 1, we have

$$\frac{1}{m} \sum_{i=1}^m \left(\frac{\partial}{\partial \lambda} \mathbf{h}(\mathbf{y}_i, \lambda) \right) (\mathbf{h}(\mathbf{y}_i, \lambda) - \mathbf{X}_i \boldsymbol{\beta})' |\mathbf{y}_m = O_p(1),$$

so that $\mathbf{K}_3^*|\mathbf{y}_m = O_p(1)$. Since \mathbf{K}_3^* is an $n_i \times n_i$ matrix, it follows that $K_3 = O_p(1)$. These observations show that the denominator in (33) is of order $O_p(1)$. Hence, one gets $\sqrt{m}(\widehat{\lambda} - \lambda)|\mathbf{y}_m = O_p(1)$, namely, $(\widehat{\lambda} - \lambda)|\mathbf{y}_m = O_p(m^{-1/2})$.

Finally, we show that $E[\widehat{\lambda} - \lambda | \mathbf{y}_m] = O_p(m^{-1})$. Evaluating the term in (33) more precisely based on the fact that $(\widehat{\lambda} - \lambda) | \mathbf{y}_m = O_p(m^{-1/2})$, we can approximate $\widehat{\lambda} - \lambda$ stochastically as

$$\widehat{\lambda} - \lambda = -\frac{F(\widehat{\beta}(\lambda), \widehat{\sigma}_v^2(\lambda), \widehat{\sigma}_e^2(\lambda), \lambda)}{\frac{\partial}{\partial \lambda} F(\widehat{\beta}(\lambda), \widehat{\sigma}_v^2(\lambda), \widehat{\sigma}_e^2(\lambda), \lambda)} + O_p(m^{-1}).$$

Let $M = E[\partial F(\widehat{\beta}(\lambda), \widehat{\sigma}_v^2(\lambda), \widehat{\sigma}_e^2(\lambda), \lambda) / \partial \lambda]$, which is of order $O(m)$. From Lemma 1 and Assumption 2, it easily follows that

$$\frac{1}{m} \left(\frac{\partial}{\partial \lambda} F(\widehat{\beta}(\lambda), \widehat{\sigma}_v^2(\lambda), \widehat{\sigma}_e^2(\lambda), \lambda) \right) = \frac{M}{m} + O_p(m^{-1/2}).$$

Then, one gets $\widehat{\lambda} - \lambda = -M^{-1} F(\widehat{\beta}(\lambda), \widehat{\sigma}_v^2(\lambda), \widehat{\sigma}_e^2(\lambda), \lambda) + O_p(m^{-1})$, so that

$$E[\widehat{\lambda} - \lambda] = -M^{-1} E[F(\widehat{\beta}(\lambda), \widehat{\sigma}_v^2(\lambda), \widehat{\sigma}_e^2(\lambda), \lambda)] + O(m^{-1}).$$

Note that $m^{-1} F(\widehat{\beta}(\lambda), \widehat{\sigma}_v^2(\lambda), \widehat{\sigma}_e^2(\lambda), \lambda)$ is evaluated as

$$\begin{aligned} m^{-1} F(\widehat{\beta}(\lambda), \widehat{\sigma}_v^2(\lambda), \widehat{\sigma}_e^2(\lambda), \lambda) &= m^{-1} F(\beta, \sigma_v^2, \sigma_e^2, \lambda) + \sigma_e^{-2} (\widehat{\beta}(\lambda) - \beta)' \frac{1}{m} \sum_{i=1}^m \mathbf{V}_i(\rho)^{-1} \left(\frac{\partial}{\partial \lambda} \mathbf{h}(\mathbf{y}_i, \lambda) \right) \\ &+ \sigma_e^{-2} \frac{1}{m} \sum_{i=1}^m (\mathbf{h}(\mathbf{y}_i, \lambda) - \mathbf{X}_i \beta)' \{ \mathbf{V}_i(\widehat{\rho}(\lambda))^{-1} - \mathbf{V}_i(\rho)^{-1} \} \left(\frac{\partial}{\partial \lambda} \mathbf{h}(\mathbf{y}_i, \lambda) \right) \\ &+ \left(\frac{1}{\widehat{\sigma}_e^2(\lambda)} - \frac{1}{\sigma_e^2} \right) \frac{1}{m} \sum_{i=1}^m (\mathbf{h}(\mathbf{y}_i, \lambda) - \mathbf{X}_i \beta)' \mathbf{V}_i(\rho)^{-1} \left(\frac{\partial}{\partial \lambda} \mathbf{h}(\mathbf{y}_i, \lambda) \right) + O_p(1). \end{aligned}$$

From Assumption (A.4), it is easy to see that $E[\widehat{\sigma}_e^{-2}(\lambda) - \sigma_e^{-2} | \mathbf{y}_m] = O_p(m^{-1})$ and $E[\mathbf{V}_i(\widehat{\rho}(\lambda))^{-1} - \mathbf{V}_i(\rho)^{-1} | \mathbf{y}_m] = O_p(m^{-1})$, which conclude that $E[m^{-1} F(\widehat{\beta}(\lambda), \widehat{\sigma}_v^2(\lambda), \widehat{\sigma}_e^2(\lambda), \lambda) | \mathbf{y}_m] = O_p(m^{-1})$. Therefore, the proof is complete. \square

A.4 Proof of Lemma 4. From Lemma 3, we need to establish the results for $\widehat{\beta}(\widehat{\lambda})$ and $\widehat{\sigma}^2(\widehat{\lambda})$. Let $i = m$. From Lemmas 1 and 3, we have

$$\begin{aligned} (\widehat{\beta}(\widehat{\lambda}) - \beta) | \mathbf{y}_m &= \left(\frac{\partial}{\partial \lambda} \widehat{\beta}(\lambda) \right) (\widehat{\lambda} - \lambda) | \mathbf{y}_m + O_p(m^{-1}) \\ &= \left(\frac{\partial}{\partial \lambda} \widehat{\beta}(\lambda) - E \left[\frac{\partial}{\partial \lambda} \widehat{\beta}(\lambda) | \mathbf{y}_m \right] \right) (\widehat{\lambda} - \lambda) | \mathbf{y}_m \\ &\quad + E \left[\frac{\partial}{\partial \lambda} \widehat{\beta}(\lambda) | \mathbf{y}_m \right] (\widehat{\lambda} - \lambda) | \mathbf{y}_m + O_p(m^{-1}) \\ &= E \left[\frac{\partial}{\partial \lambda} \widehat{\beta}(\lambda) | \mathbf{y}_m \right] (\widehat{\lambda} - \lambda) | \mathbf{y}_m + O_p(m^{-1}), \end{aligned}$$

since $E[\partial\hat{\beta}(\lambda)/\partial\lambda|\mathbf{y}_m] = \mathbf{O}_p(1)$. Then, one gets $(\hat{\beta}(\hat{\lambda}) - \beta)|\mathbf{y}_m = \mathbf{O}_p(m^{-1/2})$ and $E[\hat{\beta}(\hat{\lambda}) - \beta|\mathbf{y}_m] = \mathbf{O}_p(m^{-1})$ from Lemmas 1 and 3. Similarly, the results for $\hat{\sigma}^2(\hat{\lambda})$ follow from Lemma 3 and Assumption 2, since

$$(\hat{\sigma}^2(\hat{\lambda}) - \sigma^2)|\mathbf{y}_m = E\left[\frac{\partial}{\partial\lambda}\hat{\sigma}^2(\lambda)|\mathbf{y}_m\right](\hat{\lambda} - \lambda)|\mathbf{y}_m + \mathbf{O}_p(m^{-1}),$$

where $E[\partial\hat{\sigma}^2(\lambda)/\partial\lambda|\mathbf{y}_m] = \mathbf{O}_p(1)$. Therefore, the proof is complete. \square

A.5 Proof of Theorem 1. It is first noted that in the proof, the capital C , with or without suffix, means a generic constant. If $\mathcal{L}_m(q)$ is expanded as

$$\mathcal{L}_m(q) = \Phi(q) + m^{-1}\gamma(q, \boldsymbol{\theta}) + \mathbf{O}_p(m^{-3/2}), \quad (34)$$

where $\gamma(q, \boldsymbol{\theta})$ is a smooth function with $O(1)$, then the corresponding expansion holds for $\mathcal{L}_m^*(q)$, namely,

$$\mathcal{L}_m^*(q) = \Phi(q) + m^{-1}\gamma(q, \hat{\boldsymbol{\theta}}) + \mathbf{O}_p(m^{-3/2}).$$

Thus, one gets

$$\begin{aligned} \mathcal{L}_m^*(q) - \mathcal{L}_m(q) &= m^{-1}\{\gamma(q, \hat{\boldsymbol{\theta}}) - \gamma(q, \boldsymbol{\theta})\} + \mathbf{O}_p(m^{-3/2}) \\ &= m^{-1}\left(\frac{\partial\gamma(q, \boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right)'(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) + \mathbf{O}_p(m^{-3/2}), \end{aligned} \quad (35)$$

which establishes the result given in Theorem 1. Hence, we shall show the expansion (34) through the following steps.

(Step 1) Expansion of $\mathcal{L}_m(q)$. Since the inequality $\hat{\sigma}_i^{-1}\{h(h^{-1}(\xi_i, \lambda), \hat{\lambda}) - \hat{\xi}_i^{EB}\} \leq q$ for any $q \in \mathbb{R}$ is equivalently rewritten as $h^{-1}(\xi_i, \lambda) \leq h^{-1}(\hat{\xi}_i^{EB} + q\hat{\sigma}_i, \hat{\lambda})$, we have

$$\begin{aligned} \mathcal{L}_m(q) &= P[\hat{\sigma}_i^{-1}\{h(h^{-1}(\xi_i, \lambda), \hat{\lambda}) - \hat{\xi}_i^{EB}\} \leq q] \\ &= E\left(P[\sigma_i^{-1}(\xi_i - \hat{\xi}_i(\boldsymbol{\theta})) \leq \sigma_i^{-1}\{h(h^{-1}(\hat{\xi}_i^{EB} + q\hat{\sigma}_i, \hat{\lambda}), \lambda) - \hat{\xi}_i(\boldsymbol{\theta})\}]\middle|\mathbf{Y}\right) \\ &= E[\Phi(q + R(q, \mathbf{Y}))], \end{aligned}$$

where $\Phi(\cdot)$ is a cumulative distribution function of the standard normal distribution and

$$R(q, \mathbf{Y}) = \sigma_i^{-1}\{h(h^{-1}(\hat{\xi}_i^{EB} + q\hat{\sigma}_i, \hat{\lambda}), \lambda) - \hat{\xi}_i(\boldsymbol{\theta})\} - q.$$

For the standard normal density function $\phi(\cdot)$, the first and second derivatives are written as $\phi'(x) = -x\phi(x)$ and $\phi''(x) = (x^2 - 1)\phi(x)$ for $x \in \mathbb{R}$. The Taylor expansion is applied to get

$$\begin{aligned} \mathcal{L}_m(q) &= \Phi(q) + \phi(q)E[R(q, \mathbf{Y})] - \frac{1}{2}q\phi(q)E[R^2(q, \mathbf{Y})] \\ &\quad + \frac{1}{2}E\left[\int_q^{q+R(q, \mathbf{Y})} (q + R(q, \mathbf{Y}) - x)^2(x^2 - 1)\phi(x)dx\right] \\ &= \Phi(q) + \phi(q)t_1(q) - \frac{1}{2}q\phi(q)t_2(q) + t_3(q), \end{aligned}$$

where $t_1(q) = E[R]$, $t_2(q) = E[R^2]$ and $t_3(q) = 2^{-1}E[\int_q^{q+R}(q+R-x)^2(x^2-1)\phi(x)dx]$ for $R = R(q, \mathbf{Y})$. Note that $0 \leq |q+R-x| \leq |R|$ and $(x^2-1)\phi(x) \leq 2\phi(\sqrt{3})$ for $x \in (q, q+R)$. Then,

$$E\left[\int_q^{q+R}(q+R-x)^2(x^2-1)\phi(x)dx\right] \leq E\left[R^2 \int_q^{q+R} 2\phi(\sqrt{3})dx\right] \leq C_1 E[|R|^3],$$

which implies that

$$\mathcal{L}_m(q) = \Phi(q) + \phi(q)t_1(q) - \frac{1}{2}q\phi(q)t_2(q) + O(E[|R|^3]). \quad (36)$$

(Step 2) Expansion of $R = R(q, \mathbf{Y})$. We shall show that $R = R(q, \mathbf{Y}) = O_p(m^{-1/2})$ based on an expansion of R . It follows from this property that $\sup_{q \in \mathbb{R}} t_3(q) = O(m^{-3/2})$ and $\sup_{q \in \mathbb{R}} t_2(q) = O(m^{-1})$. Let $Q = h^{-1}(\hat{\xi}_i^{EB} + q\hat{\sigma}_i, \hat{\lambda}) - h^{-1}(\hat{\xi}_i(\boldsymbol{\theta}) + q\sigma_i, \lambda)$. Then,

$$\begin{aligned} h(h^{-1}(\hat{\xi}_i^{EB} + q\hat{\sigma}_i, \hat{\lambda}), \lambda) &= \hat{\xi}_i(\boldsymbol{\theta}) + q\sigma_i + h_x(h^{-1}(\hat{\xi}_i(\boldsymbol{\theta}) + q\sigma_i, \lambda), \lambda)Q \\ &\quad + \frac{1}{2}h_{xx}(h^{-1}(\hat{\xi}_i(\boldsymbol{\theta}) + q\sigma_i, \lambda), \lambda)Q^2 \\ &\quad + \frac{1}{2}\int_a^{a+Q}(a+Q-x)^2h_{xxx}(x, \lambda)dx, \end{aligned}$$

where $a = h^{-1}(\hat{\xi}_i(\boldsymbol{\theta}) + q\sigma_i, \lambda)$. Since $0 \leq |q+Q-x| \leq |Q|$ for $x \in (a, a+Q)$, we have

$$\left|\int_a^{a+Q}(a+Q-x)^2h_{xxx}(x, \lambda)dx\right| \leq Q^2\left|h_{xx}(a+Q, \lambda) - h_{xx}(Q, \lambda)\right|.$$

It is here noted that $Q = O_p(m^{-1/2})$, which will be shown in (Step 3) below. Then it follows from Assumption 6 that $h_x(a+Q, \lambda)$, $h_x(Q, \lambda)$ and $h_{xx}(a, \lambda)$ are $O_p(1)$. Thus,

$$h(h^{-1}(\hat{\xi}_i^{EB} + q\hat{\sigma}_i, \hat{\lambda}), \lambda) = \hat{\xi}_i(\boldsymbol{\theta}) + q\sigma_i + g(\mathbf{y}_i, \boldsymbol{\theta})Q + O_p(m^{-1}),$$

for $g(\mathbf{y}_i, \boldsymbol{\theta}) = h_x(h^{-1}(\hat{\xi}_i(\boldsymbol{\theta}) + q\sigma_i, \lambda), \lambda)$. Since $g(\mathbf{y}_i, \boldsymbol{\theta}) = O_p(1)$, it can be observed that $g(\mathbf{y}_i, \boldsymbol{\theta})Q = O_p(m^{-1/2})$, which results in

$$R(q, \mathbf{Y}) = \sigma_i^{-1}g(\mathbf{y}_i, \boldsymbol{\theta})Q + O_p(m^{-1}). \quad (37)$$

Also, the expectation of $R(q, \mathbf{Y})$ is evaluated as

$$E[R(q, \mathbf{Y})] = \sigma_i^{-1}E[g(\mathbf{y}_i, \boldsymbol{\theta})E[Q|\mathbf{y}_i]] + O(m^{-1}). \quad (38)$$

(Step 3) Evaluation of Q and $E[Q|\mathbf{y}_i]$. To get the expansion (34), it is sufficient to show that $Q = O_p(m^{-1/2})$ and $E[Q|\mathbf{y}_i] = O_p(m^{-1})$ from (38). To this end, decompose Q as $Q = Q_1 + Q_2$, where

$$Q_1 = h^{-1}(\hat{\xi}_i^{EB} + q\hat{\sigma}_i, \lambda) - h^{-1}(\hat{\xi}_i(\boldsymbol{\theta}) + q\sigma_i, \lambda), \quad (39)$$

$$Q_2 = h^{-1}(\hat{\xi}_i^{EB} + q\hat{\sigma}_i, \hat{\lambda}) - h^{-1}(\hat{\xi}_i^{EB} + q\hat{\sigma}_i, \lambda). \quad (40)$$

From (39), Q_1 is expanded as

$$Q_1 = h_x^{-1}(\hat{\xi}_i(\boldsymbol{\theta}) + q\sigma_i, \lambda)U + h_{xx}^{-1}(\hat{\xi}_i(\boldsymbol{\theta}) + q\sigma_i, \lambda)U^2 \\ + \frac{1}{2} \int_b^{b+U} (b + Q - x)^2 h_{xxx}^{-1}(x, \lambda) dx,$$

where $U = \hat{\xi}_i^{EB} - \hat{\xi}_i(\boldsymbol{\theta}) + q(\hat{\sigma}_i - \sigma_i)$ and $b = \hat{\xi}_i(\boldsymbol{\theta}) + q\sigma_i$. It is here noted that

$$U|\mathbf{y}_i = O_p(m^{-1/2}) \quad \text{and} \quad E[U|\mathbf{y}_i] = O_p(m^{-1}), \quad (41)$$

which will be shown in (Step 4) below. Then, it follows that the last two terms of the expansion of Q_1 are $O_p(m^{-1})$ given \mathbf{y}_i , and $Q_1|\mathbf{y}_i = O_p(m^{-1/2})$ by the similar argument. Thus,

$$E[Q_1|\mathbf{y}_i] = h_x^{-1}(\hat{\xi}_i(\boldsymbol{\theta}) + q\sigma_i, \lambda)E(U|\mathbf{y}_i) + O_p(m^{-1}) = O_p(m^{-1}).$$

Also, Q_2 is expanded as

$$Q_2 = h_\lambda^{-1}(\hat{\xi}_i^{EB} + q\hat{\sigma}_i, \lambda)(\hat{\lambda} - \lambda) + \frac{1}{2}h_{\lambda\lambda}^{-1}(\hat{\xi}_i^{EB} + q\hat{\sigma}_i, \lambda^*)(\hat{\lambda} - \lambda)^2,$$

where λ^* is intermediate value between λ and $\hat{\lambda}$. It can be observed that $h_\lambda^{-1}(\hat{\xi}_i^{EB} + q\hat{\sigma}_i, \lambda)|\mathbf{y}_i = O_p(1)$, $h_{\lambda\lambda}^{-1}(\hat{\xi}_i^{EB} + q\hat{\sigma}_i, \lambda^*)|\mathbf{y}_i = O_p(1)$ under Assumption 6 and $(\hat{\lambda} - \lambda)|\mathbf{y}_i = O_p(m^{-1/2})$ from Lemma 4. Thus, $Q_2|\mathbf{y}_i = O_p(m^{-1/2})$ and

$$E[Q_2|\mathbf{y}_i] = E[h_\lambda^{-1}(\hat{\xi}_i^{EB} + q\hat{\sigma}_i, \lambda)(\hat{\lambda} - \lambda)|\mathbf{y}_i] + O_p(m^{-1}) \\ = E[\{h_\lambda^{-1}(\hat{\xi}_i^{EB} + q\hat{\sigma}_i, \lambda) - h_\lambda^{-1}(\hat{\xi}_i(\boldsymbol{\theta}) + q\sigma_i, \lambda)\}(\hat{\lambda} - \lambda)|\mathbf{y}_i] \\ + h_\lambda^{-1}(\hat{\xi}_i(\boldsymbol{\theta}) + q\sigma_i, \lambda)E(\hat{\lambda} - \lambda|\mathbf{y}_i) + O_p(m^{-1}) \\ = O_p(m^{-1}),$$

since $h_\lambda^{-1}(\hat{\xi}_i^{EB} + q\hat{\sigma}_i, \lambda) - h_\lambda^{-1}(\hat{\xi}_i(\boldsymbol{\theta}) + q\sigma_i, \lambda)$ given \mathbf{y}_i is $O_p(m^{-1/2})$, which can be verified by $h_{\lambda x}^{-1}(\hat{\xi}_i(\boldsymbol{\theta}) + q\sigma_i, \lambda) = O_p(1)$ and Lemma 4.

(Step 4) Evaluation of $U|\mathbf{y}_i$ and $E[U|\mathbf{y}_i]$. It remains to show that $U|\mathbf{y}_i = O_p(m^{-1/2})$ and $E(U|\mathbf{y}_i) = O_p(m^{-1})$, for which it is sufficient to show that both $(\hat{\xi}_i^{EB} - \hat{\xi}_i(\boldsymbol{\theta}))|\mathbf{y}_i$ and $(\hat{\sigma}_i - \sigma_i)|\mathbf{y}_i$ are $O_p(m^{-1/2})$ and the conditional expectation given \mathbf{y}_i is $O_p(m^{-1})$. Recall that $U = \hat{\xi}_i^{EB} - \hat{\xi}_i(\boldsymbol{\theta}) + q(\hat{\sigma}_i - \sigma_i)$. First, $\hat{\xi}_i^{EB} - \hat{\xi}_i(\boldsymbol{\theta})$ is rewritten as

$$\hat{\xi}_i^{EB} - \hat{\xi}_i(\boldsymbol{\theta}) = \bar{\mathbf{x}}_i'(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) + \frac{\hat{\rho}n_i}{1 + \hat{\rho}n_i}(z_i(\hat{\lambda}) - z_i(\lambda) - \bar{\mathbf{x}}_i'(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})) \\ + \left(\frac{\hat{\rho}n_i}{1 + \hat{\rho}n_i} - \frac{\rho n_i}{1 + \rho n_i} \right)(z_i(\lambda) - \bar{\mathbf{x}}_i'\boldsymbol{\beta}).$$

Note that given \mathbf{y}_i , $z_i(\hat{\lambda}) - z_i(\lambda) = z_{i,\lambda}(\lambda)(\hat{\lambda} - \lambda) + O_p(m^{-1})$ and

$$\frac{\hat{\rho}n_i}{1 + \hat{\rho}n_i} = \frac{\rho n_i}{1 + \rho n_i} + \frac{n_i}{(1 + \rho n_i)^2}(\hat{\rho} - \rho) + O_p(m^{-1}).$$

Further, from Lemma 4 and a similar expansion as in (31), it follows that $(\widehat{\rho} - \rho)|\mathbf{y}_i = O_p(m^{-1/2})$ and $E(\widehat{\rho} - \rho|\mathbf{y}_i) = O_p(m^{-1})$. Hence, one gets $(\widehat{\xi}_i^{EB} - \widehat{\xi}_i(\boldsymbol{\theta}))|\mathbf{y}_i = O_p(m^{-1/2})$ and

$$\begin{aligned} E[\widehat{\xi}_i^{EB} - \widehat{\xi}_i(\boldsymbol{\theta})|\mathbf{y}_i] &= E\left[\frac{\rho n_i}{1 + \rho n_i} z_{i,\lambda}(\lambda)(\widehat{\lambda} - \lambda) + \frac{n_i}{(1 + \rho n_i)^2} (\widehat{\rho} - \rho)(z_i(\lambda) - \bar{\mathbf{x}}_i' \boldsymbol{\beta}) \middle| \mathbf{y}_i\right] + O_p(m^{-1}) \\ &= \frac{\rho n_i}{1 + \rho n_i} z_{i,\lambda}(\lambda) E[\widehat{\lambda} - \lambda|\mathbf{y}_i] + \frac{n_i}{(1 + \rho n_i)^2} (z_i(\lambda) - \bar{\mathbf{x}}_i' \boldsymbol{\beta}) E[\widehat{\rho} - \rho|\mathbf{y}_i] + O_p(m^{-1}), \end{aligned}$$

which is of order $O_p(m^{-1})$ from Lemma 4. A similar evaluation for $\widehat{\sigma}_i - \sigma_i$ shows that given \mathbf{y}_i ,

$$\widehat{\sigma}_i - \sigma_i = \frac{1}{2} \sigma_v^{-1} (1 + n_i \rho)^{-1/2} (\widehat{\sigma}_v^2 - \sigma_v^2) - \frac{n_i}{2} \sigma_i^{-3} (\widehat{\rho} - \rho) + O_p(m^{-1}).$$

Then, from Lemma 4, it follows that $(\widehat{\sigma}_i - \sigma_i)|\mathbf{y}_i = O_p(m^{-1/2})$ and $E[\widehat{\sigma}_i - \sigma_i|\mathbf{y}_i] = O_p(m^{-1})$, which completes the proof. \square

A.6 Proof of Theorem 2. As in the proof of Theorem 1, we obtain an asymptotic expansion of $\mathcal{L}_m^c(q|\mathbf{y}_i)$ in the same settings of the proof of Theorem 1. Then for any $q \in \mathbb{R}$, we have

$$\mathcal{L}_m^c(q|\mathbf{y}_i) = E[\Phi(q + R(q, \mathbf{Y}))|\mathbf{y}_i].$$

Since $E[R(q, \mathbf{Y})|\mathbf{y}_i] = O_p(m^{-1})$, we have an asymptotic expansion of $\mathcal{L}_m^c(q|\mathbf{y}_i)$ as

$$\mathcal{L}_m^c(q|\mathbf{y}_i) = \Phi(q) + m^{-1} \eta(q, \boldsymbol{\theta}, \mathbf{y}_i) + O_p(m^{-3/2})$$

for an $O(1)$ smooth quantity $\eta(q, \boldsymbol{\theta}, \mathbf{y}_i)$, which leads to the result by Lemma 4. \square

References

- [1] Basu, R., Ghosh, J. K., and Mukerjee, R. (2003). Empirical Bayes prediction intervals in a normal regression model: Higher order asymptotics. *Statist. Probab. Lett.*, **63**, 197-203.
- [2] Battese, G.E., Harter, R.M. and Fuller, W.A. (1988). An error-components model for prediction of county crop areas using survey and satellite data. *J. Amer. Statist. Assoc.*, **83**, 28-36.
- [3] Booth, J. G. and Hobert, J. P. (1998). Standard errors of prediction in generalized linear mixed models. *J. Amer. Statist. Assoc.*, **93**, 262-272.
- [4] Box, G.E.P. and Cox, D.R. (1964). An analysis of transformation (with discussion). *J. R. Stat. Soc. Ser. B Stat. Methodol.*, **26**, 211-252.
- [5] Chatterjee, S., Lahiri, P. and Li, H. (2008). Parametric bootstrap approximation to the distribution of EBLUP and related predictions intervals in linear mixed models. *Ann. Statist.*, **36**, 1221-1245.
- [6] Das, K., Jiang, J. and Rao, J.N.K. (2004). Mean squared error of empirical predictor. *Ann. Statist.*, **32**, 818-840.

- [7] Datta, G. S., Ghosh, M. (2012). Small area shrinkage estimation. *Statist. Sci.*, **27**, 95-114
- [8] Datta, G. S., Ghosh, M., Smith, D. D. and Lahiri, P. (2002). On an asymptotic theory of conditional and unconditional coverage probabilities of empirical Bayes confidence intervals. *Scand. J. Stat.*, **29**, 139-152.
- [9] Datta, G.S., Kubokawa, T., Molina, I., and Rao, J.N.K. (2011). Estimation of mean squared error of model-based small area estimators. *Test*, **20**, 367-388.
- [10] Datta, G. S. and Lahiri, P. (2000). A unified measure of uncertainty of estimated best linear unbiased predictors in small area estimation problems. *Statist. Sinica*, **10**, 613-627.
- [11] Datta, G.S., Rao, J.N.K. and Smith, D.D. (2005). On measuring the variability of small area estimators under a basic area level model. *Biometrika*, **92**, 183-196.
- [12] Diao, L., Smith, D. D., Datta, G. S., Maiti, T. and Opsomer, J. D. (2014). Accurate confidence interval estimation of small area parameters under the Fay–Herriot model. *Scand. J. Stat.*, **41**, 497-515.
- [13] Fay, R. and Herriot, R. (1979). Estimators of income for small area places: an application of James–Stein procedures to census. *J. Amer. Statist. Assoc.*, **74**, 341-353.
- [14] Ghosh, M. and Rao, J.N.K. (1994). Small area estimation: An appraisal. *Statist. Sci.*, **9**, 55-93.
- [15] Hall, P. and Maiti, T. (2006a). Nonparametric estimation of mean-squared prediction error in nested-error regression models. *Ann. Statist.*, **34**, 1733-1750.
- [16] Hall, P. and Maiti, T. (2006b). On parametric bootstrap methods for small area prediction. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, **68**, 221-238.
- [17] Jiang, J. (2010). *Large Sample Techniques for Statistics*. Springer.
- [18] Jiang, J and Nguyen, T. (2012). Small area estimation via heteroscedastic nested-error regression, *Canad. J. Statist.*, **40**, 588-603.
- [19] Li, H. and Lahiri, P. (2010). An adjusted maximum likelihood method for solving small area estimation problems. *J. Multivariate Anal.*, **101**, 882-892.
- [20] Pfeiffermann, D. (2013) New Important Developments in Small Area Estimation. *Statist. Sci.*, **28**, 40-68
- [21] Prasad, N. and Rao, J. N. K. (1990). The estimation of mean-squared errors of small-area estimators. *J. Amer. Statist. Assoc.*, **90**, 758-766.
- [22] Rao, J.N.K. (2003). *Small Area Estimation*. Wiley.
- [23] Slud, E.V. and Maiti, T. (2006). Mean-squared error estimation in transformed Fay–Herriot models. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, **68**, 239-257.

- [24] Sugasawa, S. and Kubokawa, T. (2014). Parametric transformed Fay-Herriot model for small area estimation. *arXiv:1410.4778*.
- [25] Torabi, M. and Rao, J.N.K. (2013). Estimation of mean squared error of model-based estimators of small area means under a nested error linear regression model. *J. Multivariate Anal.*, **117**, 76 - 87.
- [26] Yang, Z. L. (2006). A modified family of power transformations. *Econ. Lett.*, **92**, 14-19.
- [27] Yoshimori, M. and Lahiri, P. (2014). A New Adjusted Maximum Likelihood Method for the Fay-Herriot Small Area Model. *J. Multivariate Anal.*, **124**, 281-294.
- [28] Yoshimori, M. and Lahiri, P. (2014). A second-order efficient empirical Bayes confidence interval. *Ann. Statist.*, **42**, 1233-1261